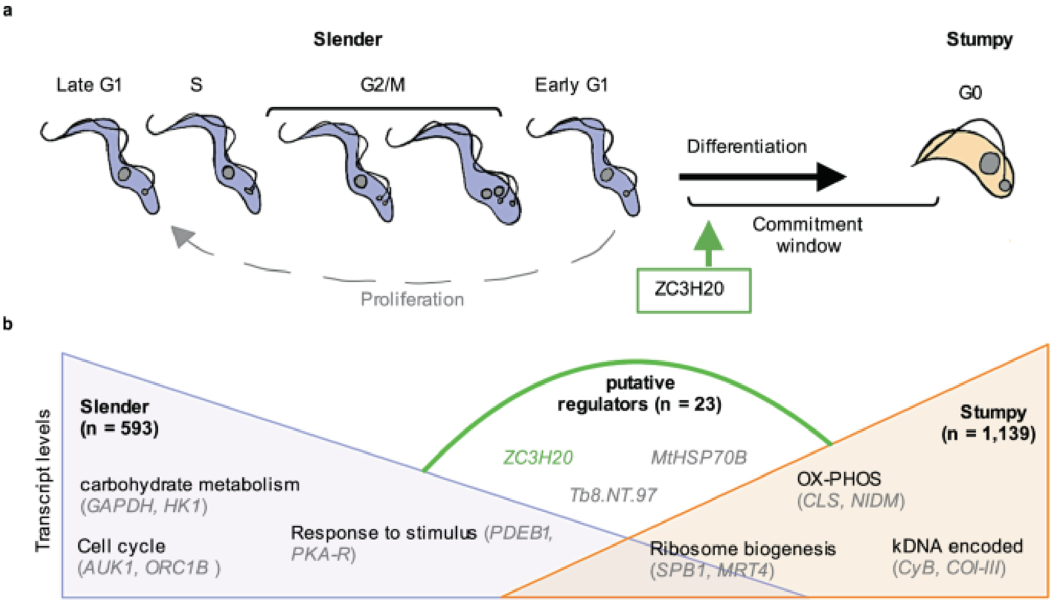
Single Cell RNA-Sequencing (scRNA-seq)

***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:

1. Find all genes with data from scRNA-seq experiments.
2. Explore scRNA-seq data on specific gene pages.
3. 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

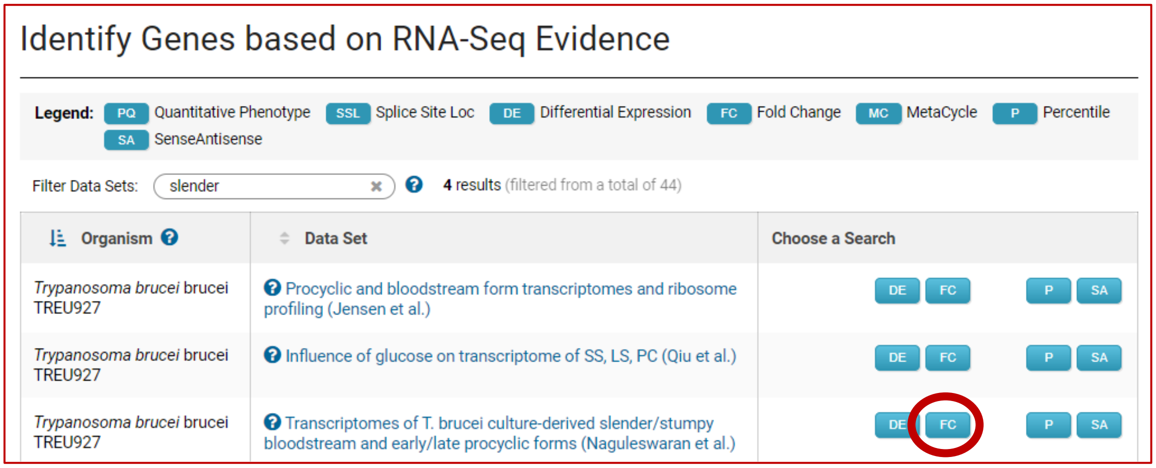
Slender markers:
GAPDH: Tb927.6.4280
PYK1: Tb927.10.14140

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. Select the fold-change search associated with the experiment ‘Transcriptomes of T. brucei culture-derived slender/stumpy bloodstream and early/late procyclic forms (Naguleswaran et al.)’ You can filter the RNA-Seq data set list with the word ‘slender’.



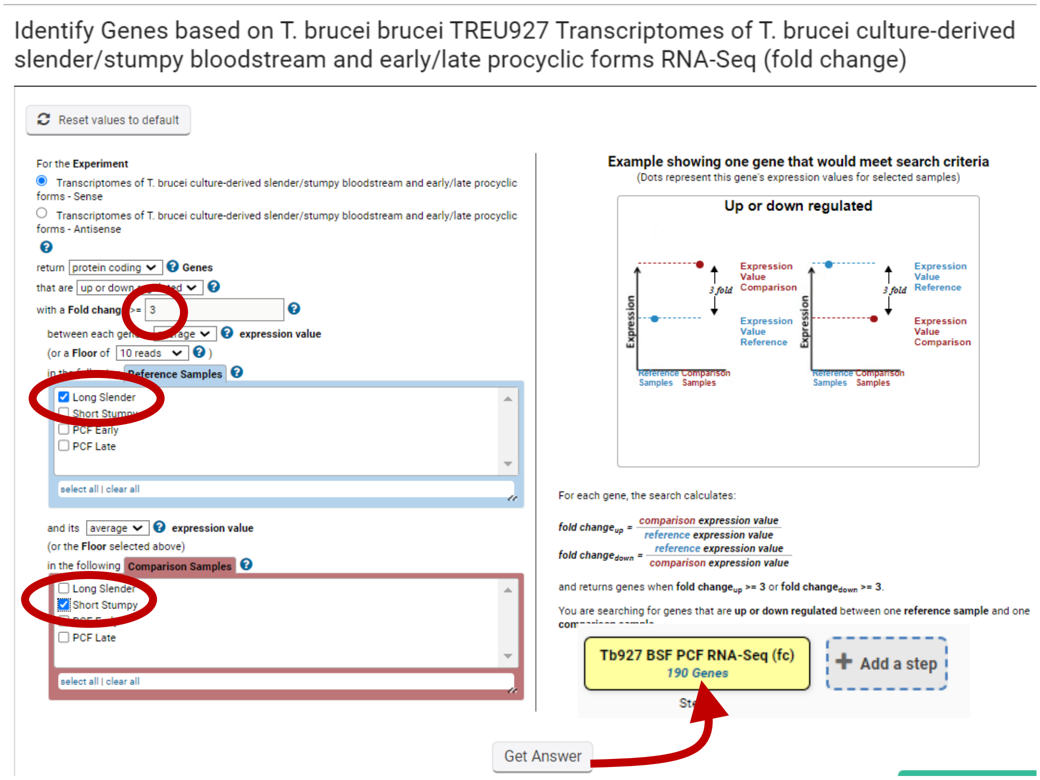
Identify Genes based on RNA-Seq Evidence

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 44)

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

b. 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.



Identify Genes based on T. brucei brucei TREU927 Transcriptomes of T. brucei culture-derived slender/stumpy bloodstream and early/late procyclic forms RNA-Seq (fold change)

Reset values to default

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 ?

with a **Fold chang** >= 3 ?

between each gen ?age ? **expression value**

(or a **Floor** of 10 reads ?)

in the f 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 Late

select all | clear all

Get Answer

Example showing one gene that would meet search criteria

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

Up or down regulated

Expression
3 fold
Expression Value Comparison
Expression Value Reference
Reference Samples
Comparison Samples
Expression
3 fold
Expression Value Reference
Expression Value Comparison
Reference Samples
Comparison Samples

For each gene, the search calculates:

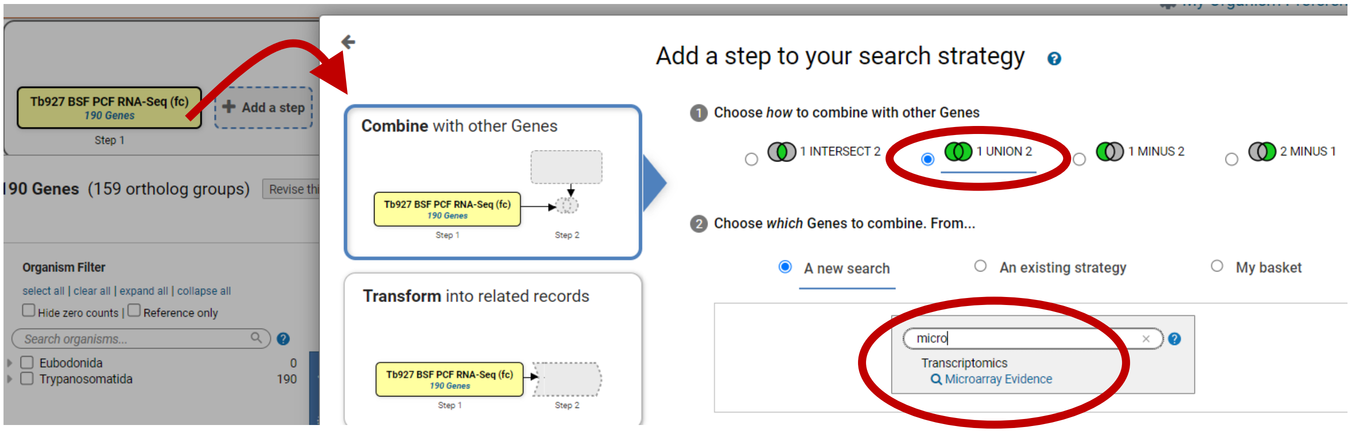
$$\textit{fold change}_{up} = \frac{\textit{comparison expression value}}{\textit{reference expression value}}$$

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

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 **com**

c. Expand your list of genes that are differentially expressed between slender and stumpy stage by searching the microarray experiment ‘Life cycle stages and differentiation time course (Kabani et al.)’.



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

Add a step

190 Genes (159 ortholog groups) Revise th

Organism Filter

select all | clear all | expand all | collapse all

☐ Hide zero counts | ☐ Reference only

Search organisms...

☐ Eubodonida	0
☐ Trypanosomatida	190

Add a step to your search strategy

Combine with other Genes

Transform into related records

1 Choose how to combine with other Genes

○ 1 INTERSECT 2 ● 1 UNION 2 ○ 1 MINUS 2 ○ 2 MINUS 1

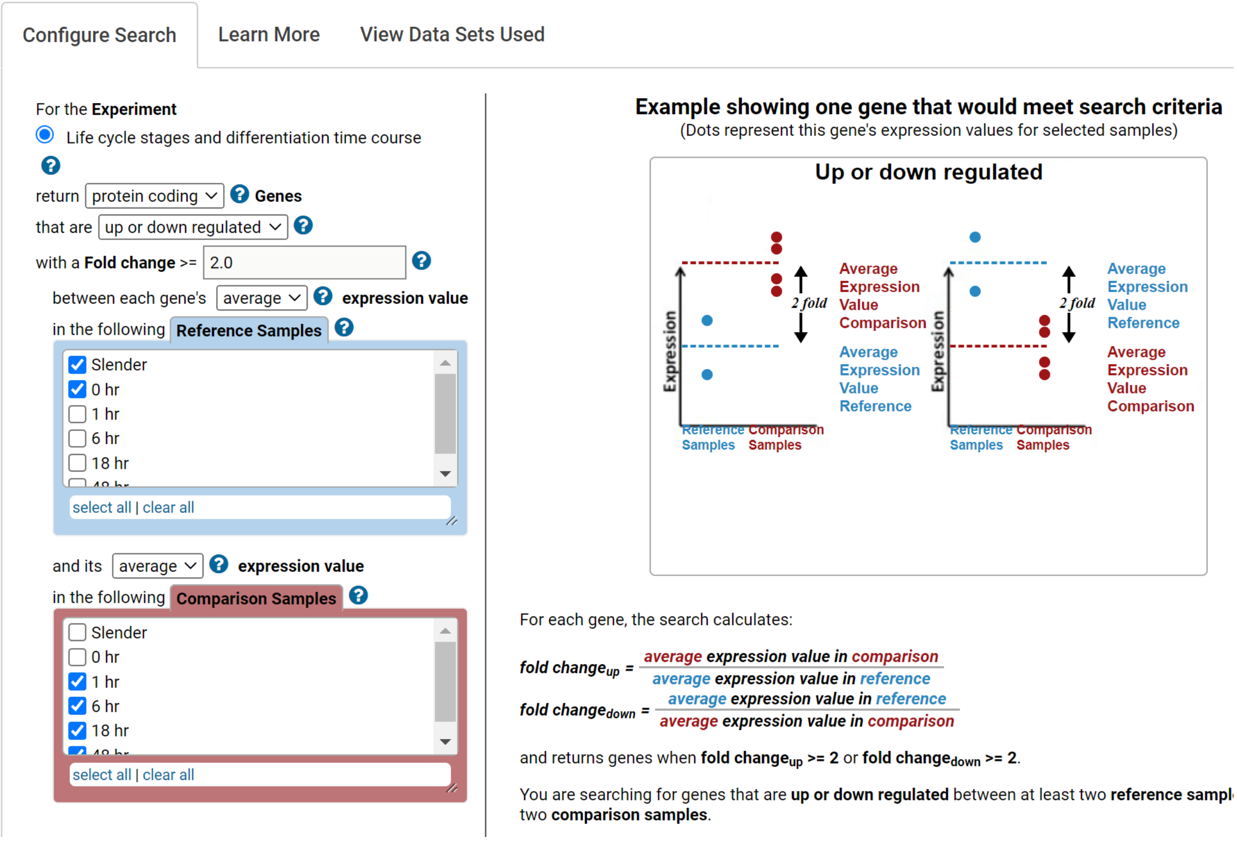
2 Choose which Genes to combine. From...

● A new search ○ An existing strategy ○ My basket

micro

Transcriptomics
Microarray Evidence

d. 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”.



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
☐ 48 hr

select all | clear all

and its average **expression value**

in the following **Comparison Samples**

☐ Slender
☐ 0 hr
☑ 1 hr
☑ 6 hr
☑ 18 hr
☑ 48 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

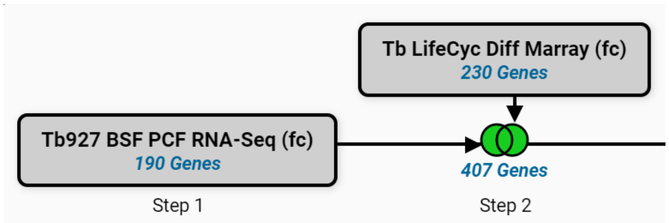
For each gene, the search calculates:

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

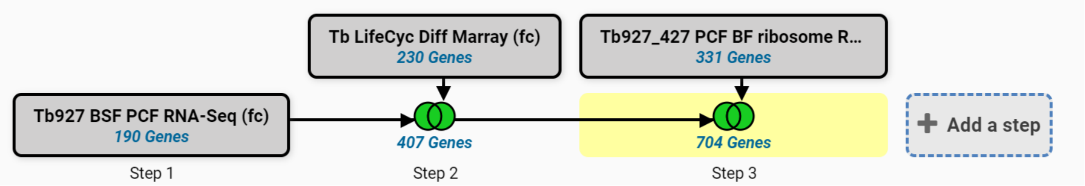
$$\textit{fold change}_{down} = \frac{\textit{average expression value in reference}}{\textit{average expression value in comparison}}$$

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**.

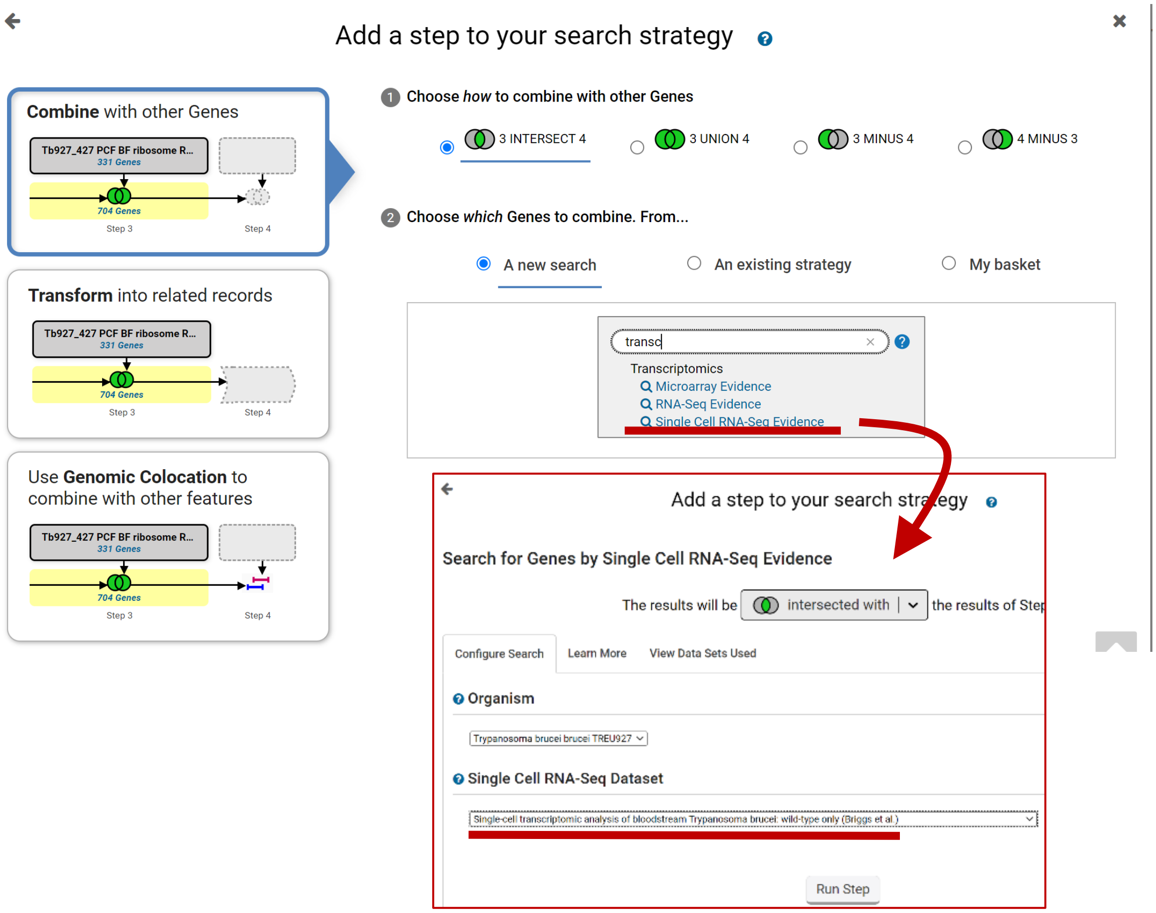


e. Using the same logic as above, add another step and find the RNA-Seq experiment from 'Procyclic and bloodstream form transcriptomes and ribosome profiling (Jensen et al.)' and configure the fold change search to find all differentially expressed genes by 2-fold comparing the blood form (**cBF mRNA**) to the slender form (**slBF mRNA**). To check your answer, here is a link to the completed strategy: https://tritrypdb.org/tritrypdb/app/workspace/strategies/import/f8edd96ff8b948e9


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. **Which stumpy/slender differentially expressed genes also have data in single cell RNA-Seq experiments.**

a. Start with the strategy from #1 above. https://tritrypdb.org/tritrypdb/app/workspace/strategies/import/f8edd96ff8b948e9 Add a step to the strategy and run the Search for Genes by Single Cell RNA-Seq Evidence. Set the Single Cell RNA-Seq Dataset parameter to "Single-cell transcriptomic analysis of bloodstream Trypanosoma brucei: **wild-type only** (Briggs et al.)". 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

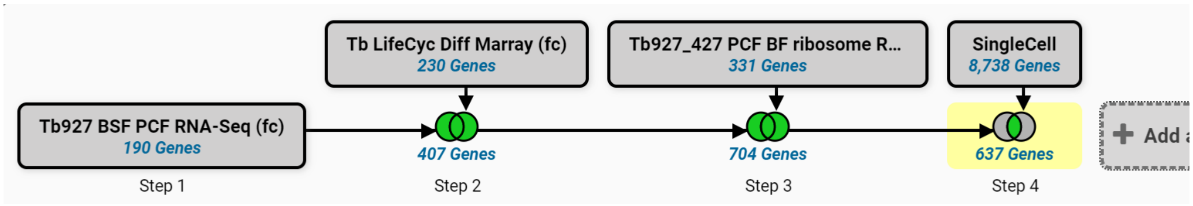
The results will be intersected with the results of Step

Configure Search | Learn More | View Data Sets Used

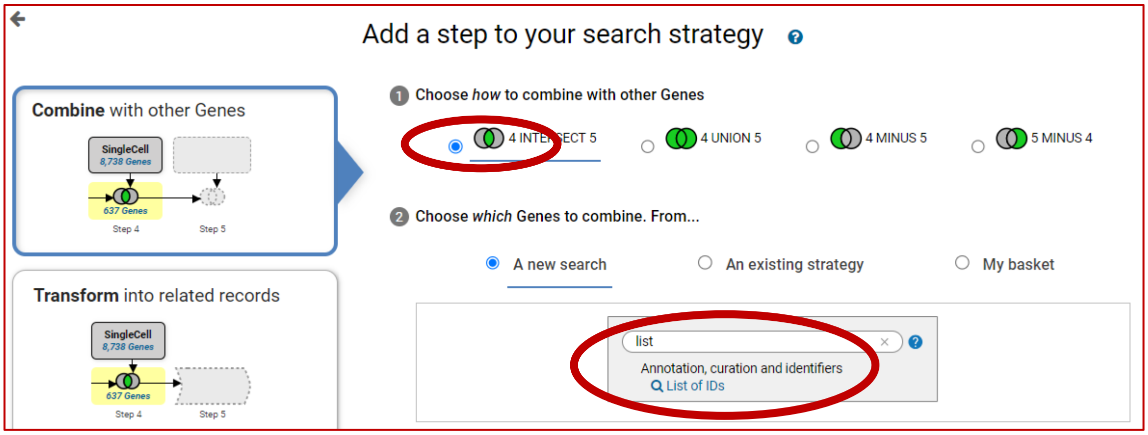
Organism: Trypanosoma brucei brucei TREU927

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

Run Step



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



Add a step to your search strategy

Combine with other Genes

Transform into related records

1 Choose *how* to combine with other Genes

● 4 INTERSECT 5 ○ 4 UNION 5 ○ 4 MINUS 5 ○ 5 MINUS 4

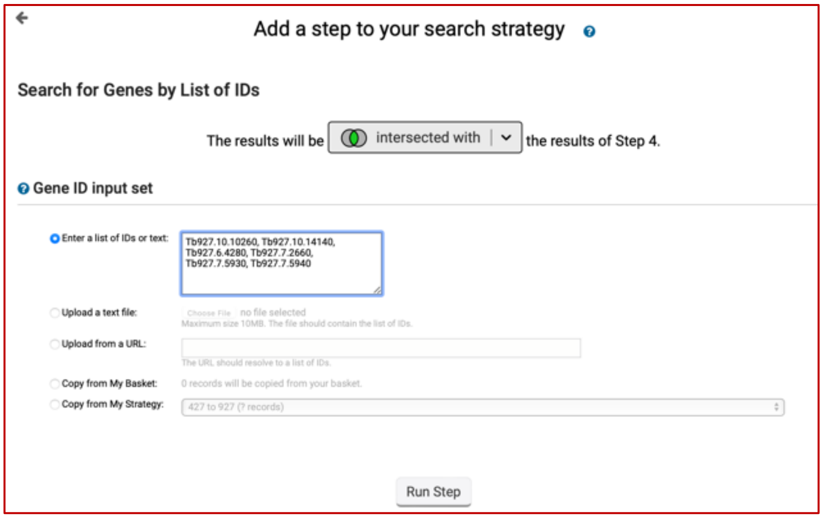
2 Choose *which* Genes to combine. From...

● A new search ○ An existing strategy ○ My basket

list

Annotation, curation and identifiers

List of IDs



Add a step to your search strategy

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

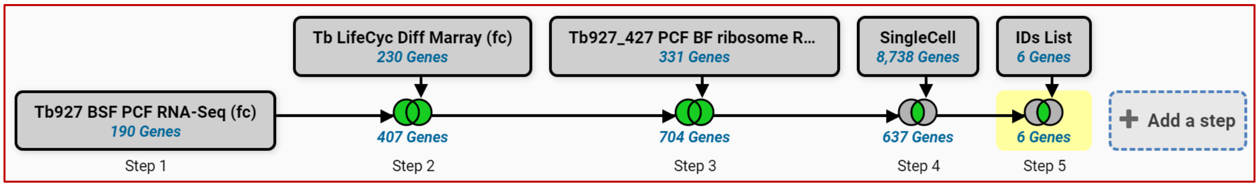
○ Upload a text file: Choose File no file selected
Maximum size 10MB. The file should contain the list of IDs.

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

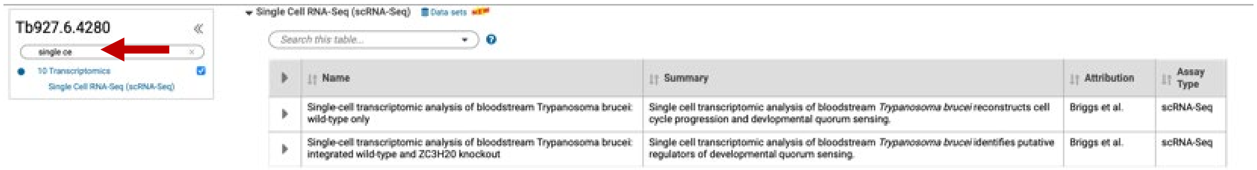
○ Copy from My Basket: 0 records will be copied from your basket.

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

Run Step



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.



Tb927.6.4280

single ce

10 Transcriptomics

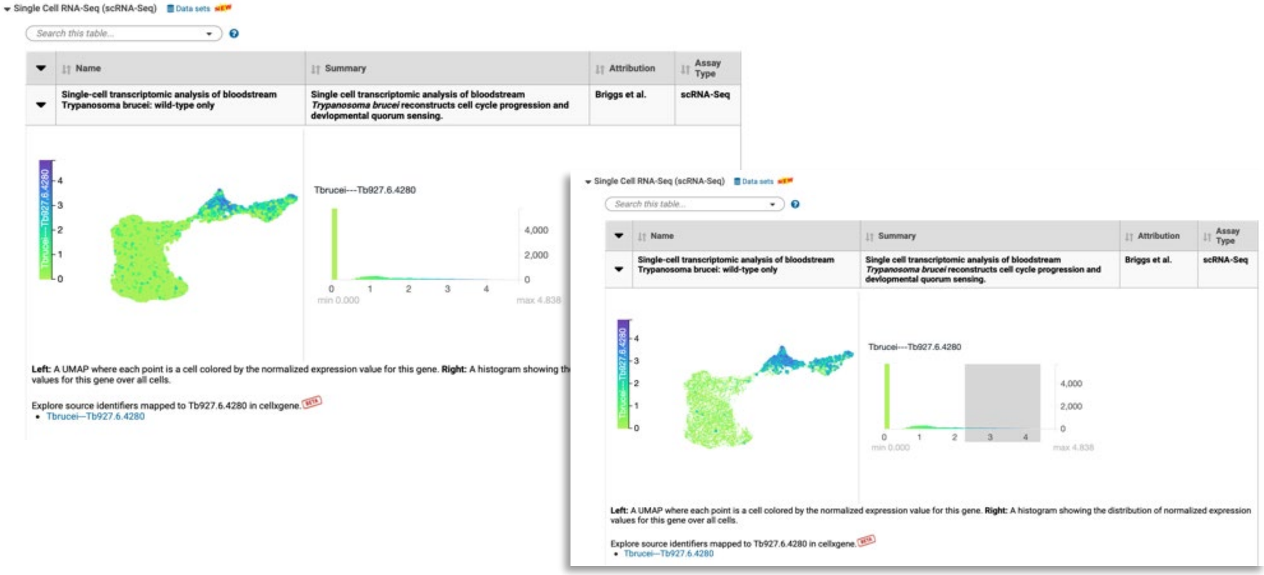
Single Cell RNA-Seq (scRNA-Seq)

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

Search this table...

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
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

Search this table...

	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 th values for this gene over all cells.

Explore source identifiers mapped to Tb927.6.4280 in cellxgene. BETA

- Tbrucei--Tb927.6.4280

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

Search this table...

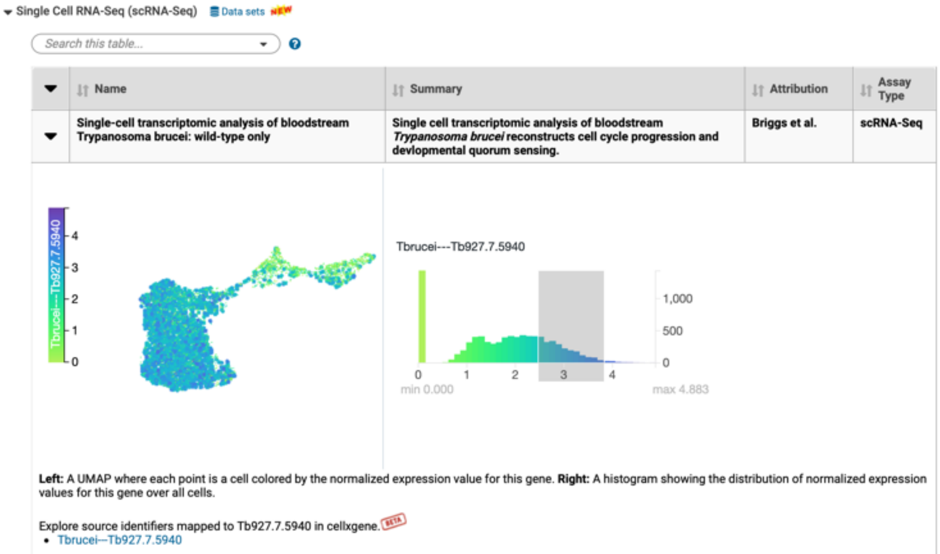
	Name	Summary	Attribution	Assay Type
▼	Single-cell transcriptomic analysis of bloodstream Trypanosoma brucei: wild-type only	Single cell transcriptomic analysis of bloodstream *Trypanosoma brucei* reconstructs cell cycle progression and devlopmental quorum sensing.	Briggs et al.	scRNA-Seq

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. BETA

- 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

Search this table...

	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.

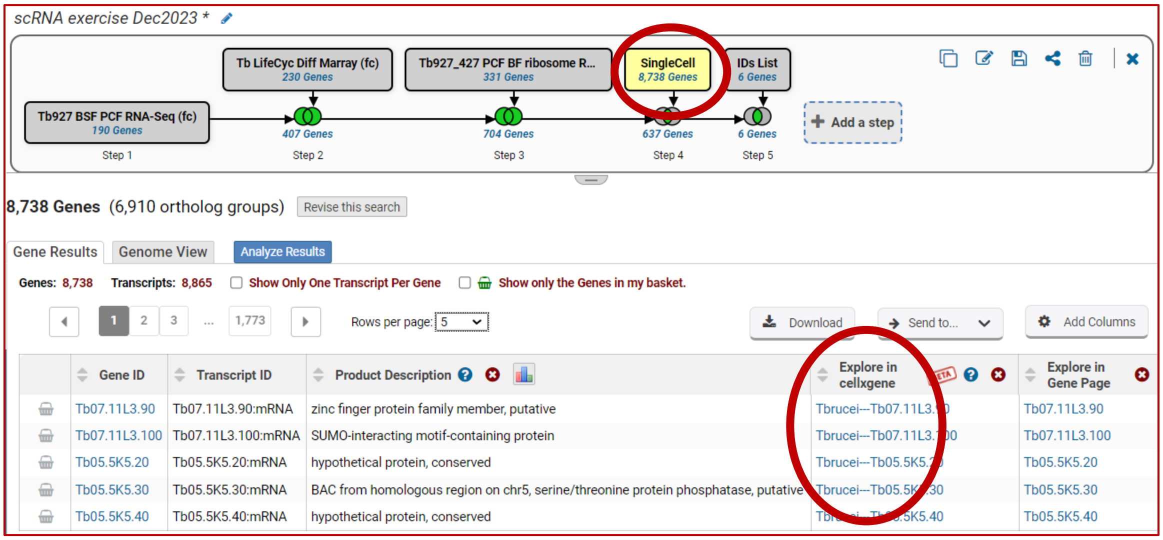
Explore source identifiers mapped to Tb927.7.5940 in cellxgene. BETA

- 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. Focus the strategy on the Single Cell result and use the ‘Explore in cellxgene’ column to open the application. (You can also reach the cellxgene application from the gene page using the link below the graphs for each experiment).



8,738 Genes (6,910 ortholog groups)

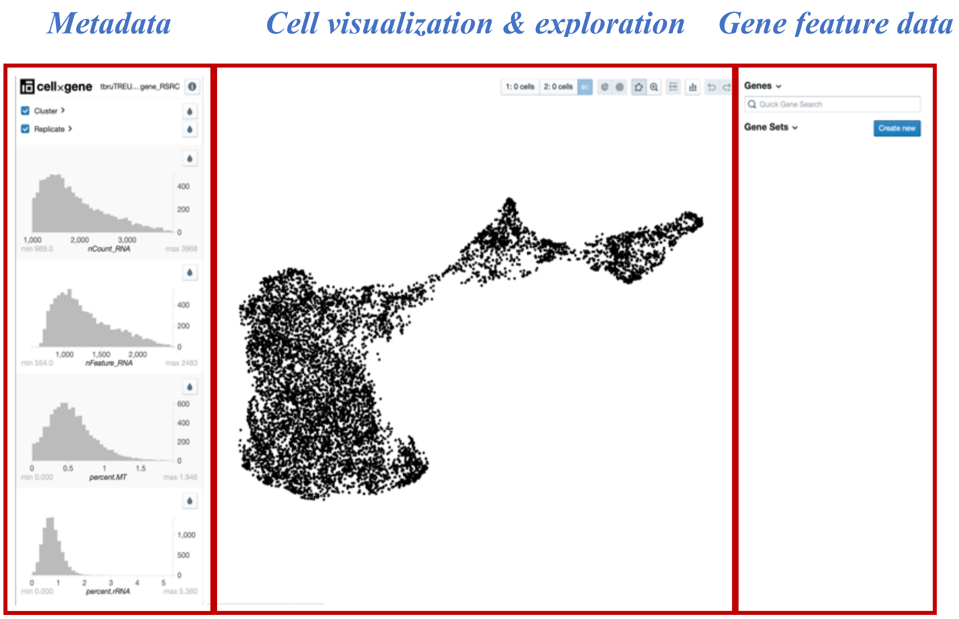
Genes: 8,738 **Transcripts:** 8,865 □ Show Only One Transcript Per Gene □ Show only the Genes in my basket.

1 2 3 ... 1,773 Rows per page: 5

	Gene ID	Transcript ID	Product Description	Explore in cellxgene	Explore in Gene Page
	Tb07.11L3.90	Tb07.11L3.90:mRNA	zinc finger protein family member, putative	Tbrucei---Tb07.11L3.90	Tb07.11L3.90
	Tb07.11L3.100	Tb07.11L3.100:mRNA	SUMO-interacting motif-containing protein	Tbrucei---Tb07.11L3.100	Tb07.11L3.100
	Tb05.5K5.20	Tb05.5K5.20:mRNA	hypothetical protein, conserved	Tbrucei---Tb05.5K5.20	Tb05.5K5.20
	Tb05.5K5.30	Tb05.5K5.30:mRNA	BAC from homologous region on chr5, serine/threonine protein phosphatase, putative	Tbrucei---Tb05.5K5.30	Tb05.5K5.30
	Tb05.5K5.40	Tb05.5K5.40:mRNA	hypothetical protein, conserved	Tbrucei---Tb05.5K5.40	Tb05.5K5.40

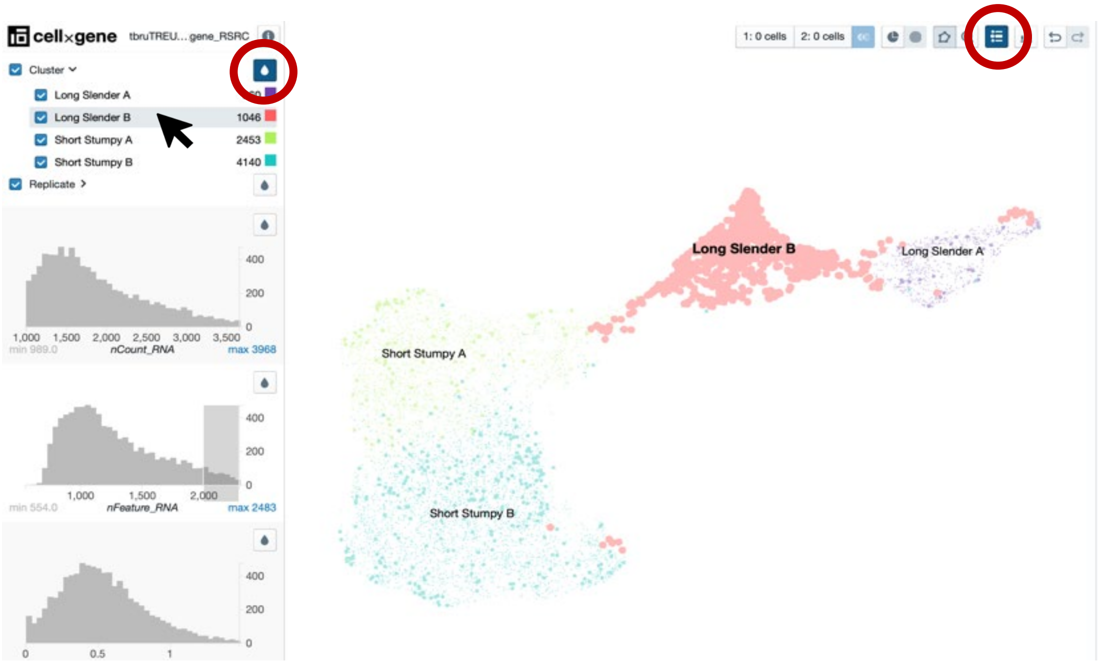
b. Your initial view will be a UMAP plot of all cells from this experiment. This may be black and white, or may be colored to show expression of a specific gene depending on how you got there.

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. 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.

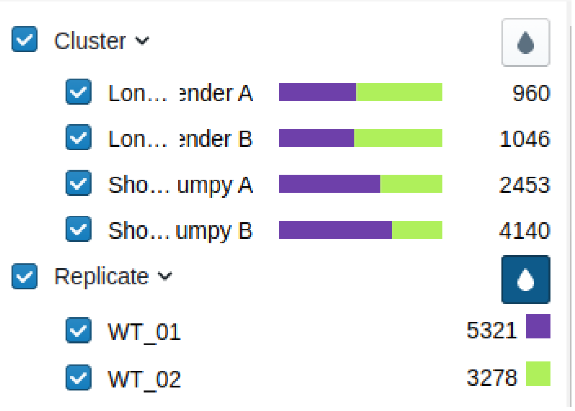


d. The droplet icon can be used to color the cells in the central panel with metadata from the left panel or gene expression data from the right panel. Try this:

- 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.
-



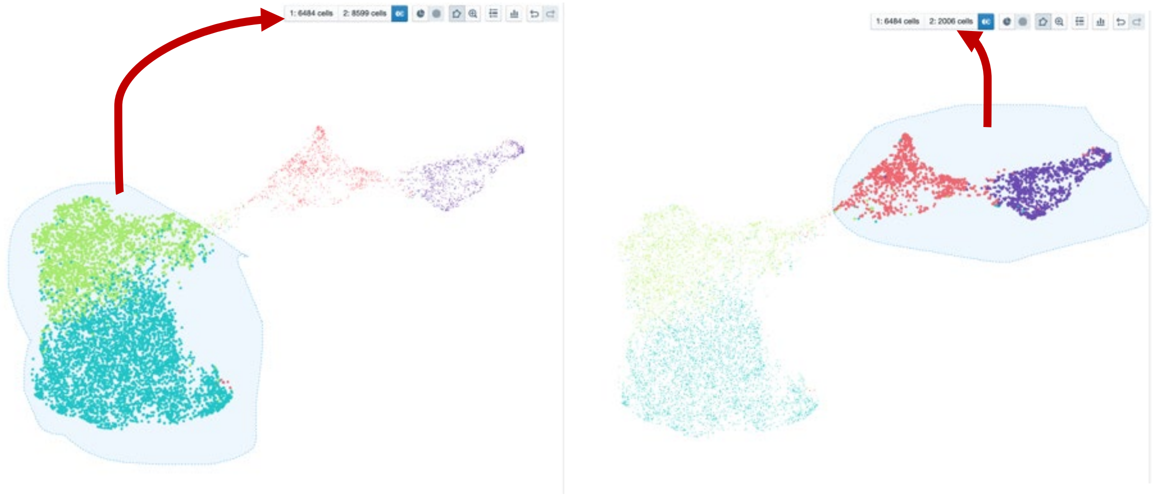
- Expand the “Replicate” metadata category. Use the droplet icon to color based on replicate. Mouseover the replicates to see how they are distributed in the UMAP. Notice the bars that appear for the cluster categories showing the proportion of cells from each replicate in each cluster. Do these look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

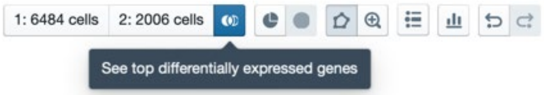


☑ Cluster ⌄	💧
☑ Lon… ender A	960
☑ Lon… ender B	1046
☑ Sho… umpy A	2453
☑ Sho… umpy B	4140
☑ Replicate ⌄	💧
☑ WT_01	5321 🟪
☑ WT_02	3278 🟩

e. 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.





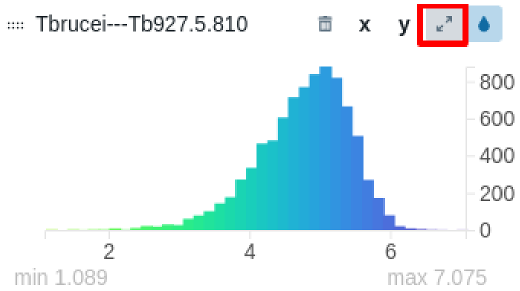
1: 6484 cells | 2: 2006 cells

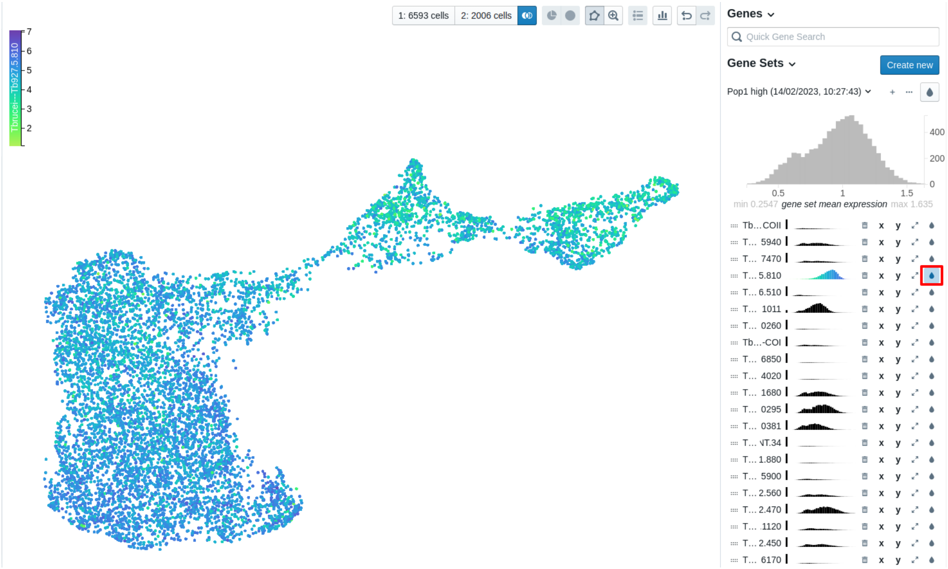
See top differentially expressed genes

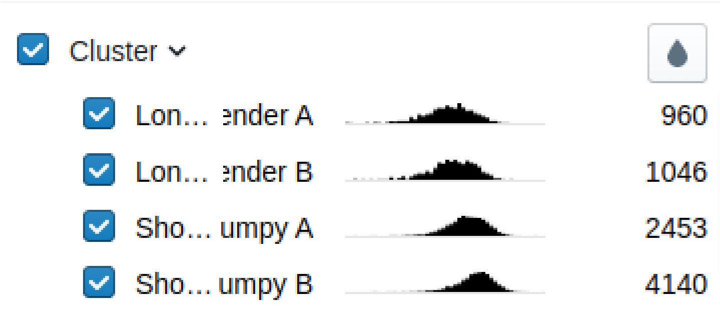
- When done with your selections and saving populations, click on the differential expression icon.

- 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.





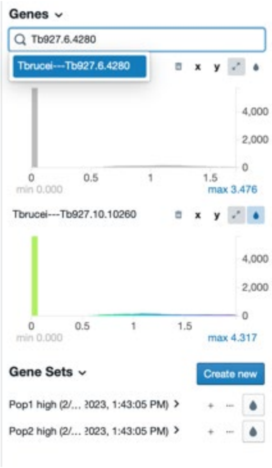


☑ Cluster	
☑ Lon… ender A	960
☑ Lon… ender B	1046
☑ Sho… umpy A	2453
☑ Sho… umpy B	4140

- 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.

f. 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

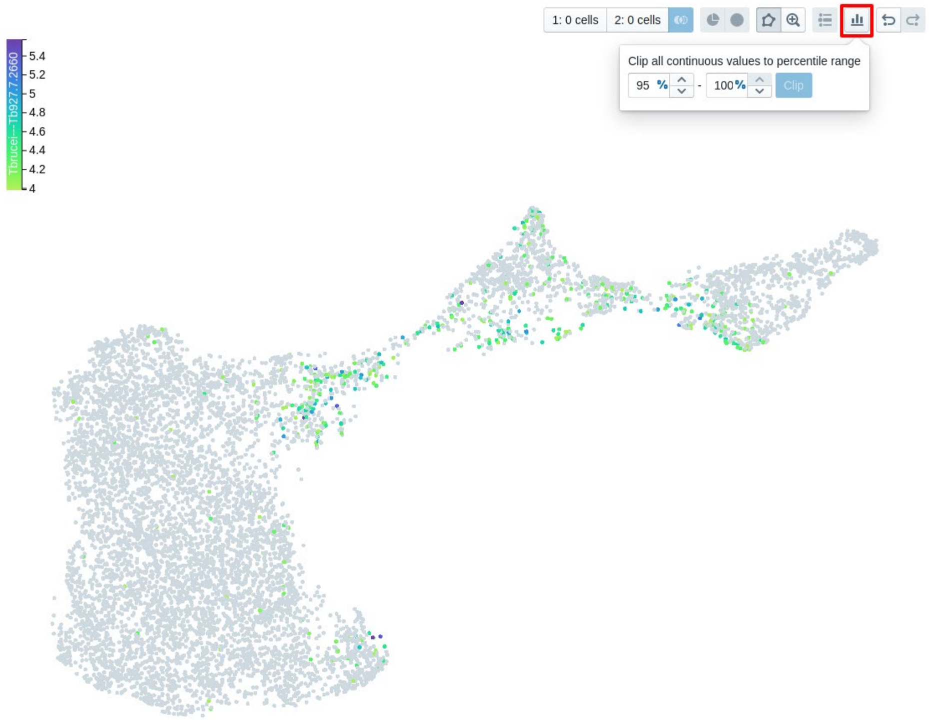


g. The authors identified one gene as a putative regulator of slender to stump transition. This is a zinc-finger protein which has been described as having a role post-transcriptional regulation. Let's look at the expression of this protein.

- 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.

- What happens to the UMAP and the histogram? Is it easier to find the cells with the highest expression levels for this gene now?



Clip all continuous values to percentile range

95 % - 100 %

- Looking at the expression levels, why do you think the authors chose this transcriptional regulator for further study?