# DEVELOPING TOOLS FOR ANNOTATION AND SINGLE CELL TRANSCRIPTOMICS ANALYSIS

DR WILLIAM HAESE-HILL

WCIP ISAB/RETREAT

31ST MAY 2022

University of Glasgow

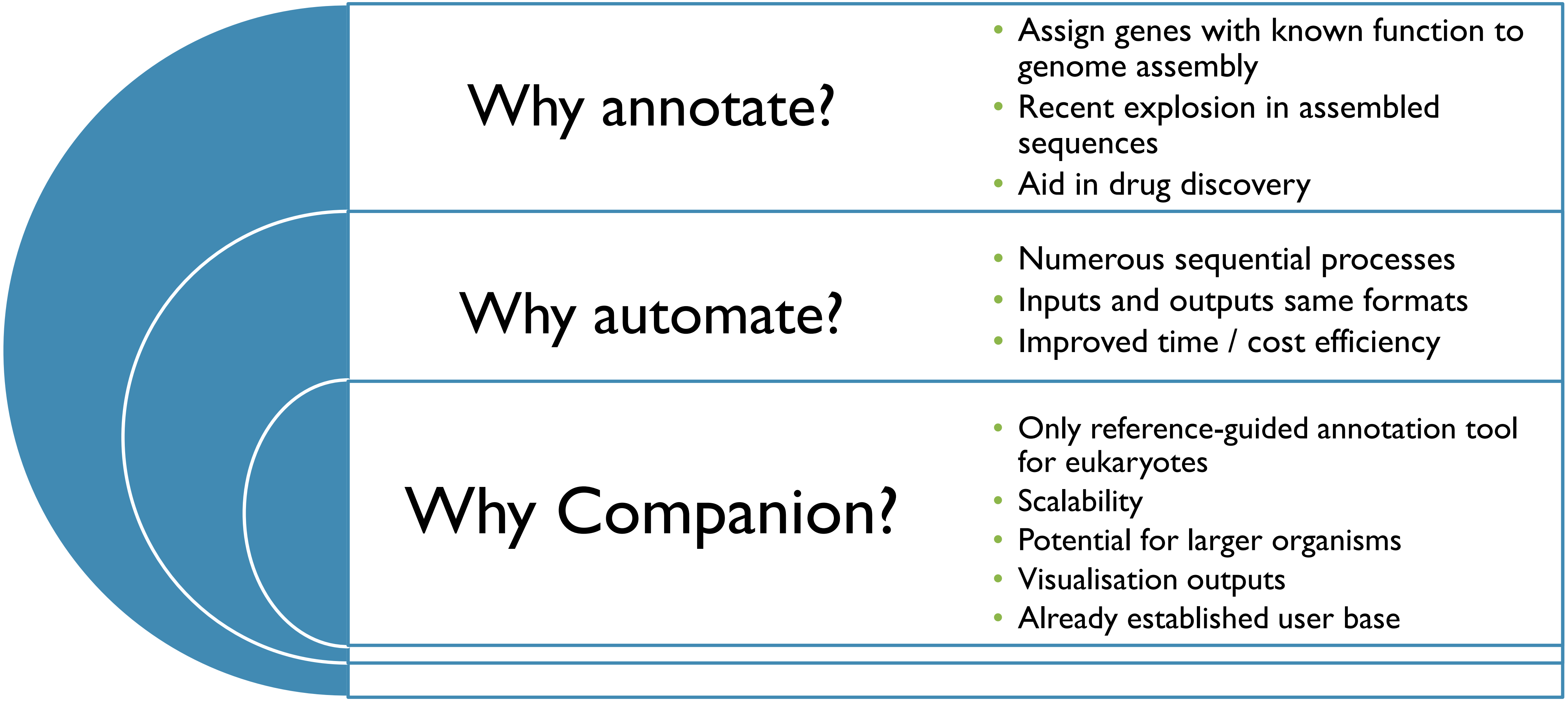
Institute of Infection, Immunity & Inflammation

wellcome centre integrative parasitology

# COMPANION

A GENOME ANNOTATION TOOL FOR MORE THAN JUST PROTISTS

# MOTIVATION



## Why annotate?

- Assign genes with known function to genome assembly
- Recent explosion in assembled sequences
- Aid in drug discovery

## Why automate?

- Numerous sequential processes
- Inputs and outputs same formats
- Improved time / cost efficiency

## Why Companion?

- Only reference-guided annotation tool for eukaryotes
- Scalability
- Potential for larger organisms
- Visualisation outputs
- Already established user base

# PIPELINE

### Pseudochromosome contiguation

- ABACAS

### Structural annotation

- RATT
- SNAP
- AUGUSTUS

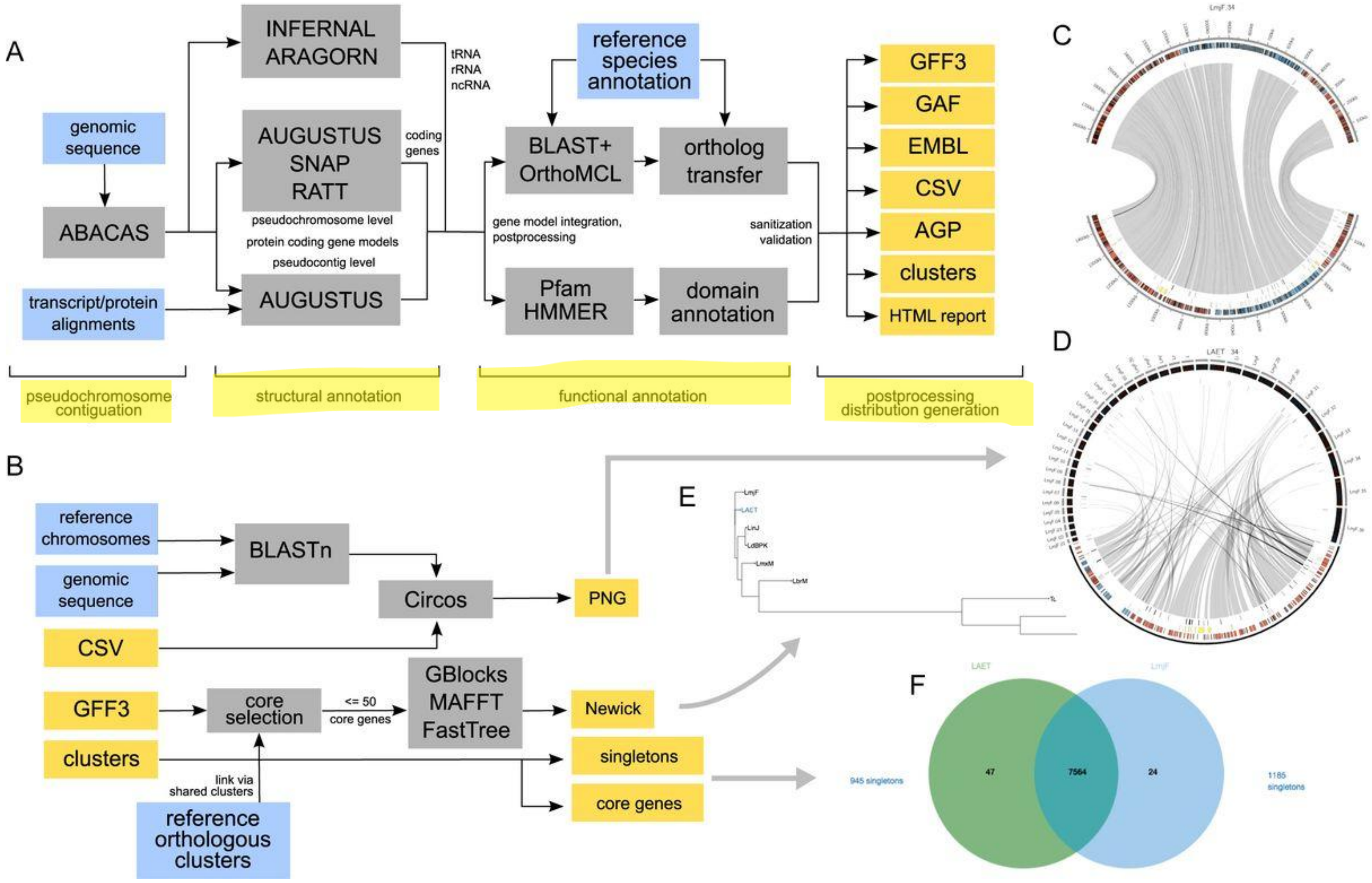
### Functional annotation

- BLAST
- OrthoMCL
- Pfam

### Outputs

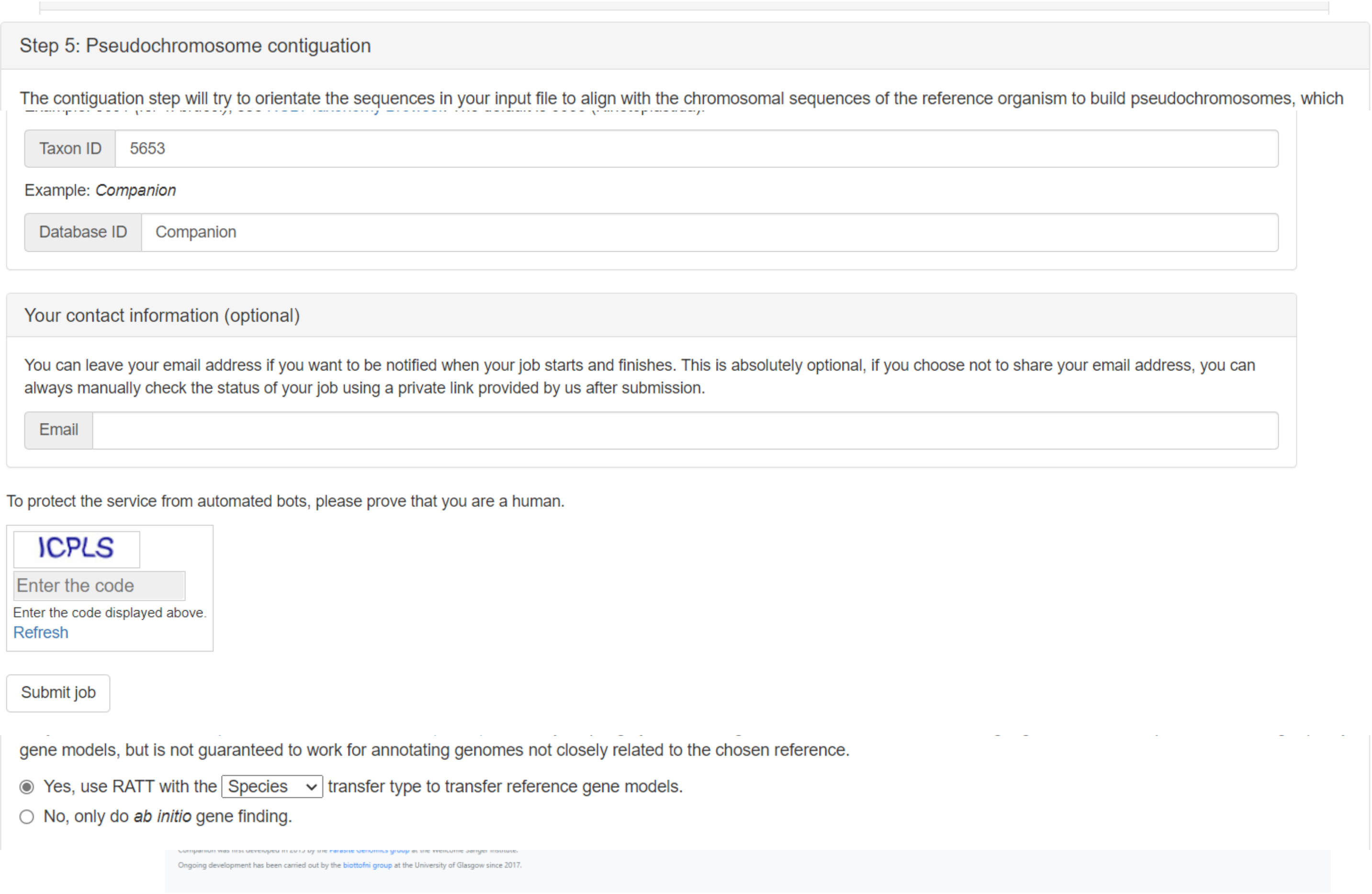
Files: GFF3, EMBL, GAF

Visuals: Orthology, Phylogeny trees, Synteny



# WEB INTERFACE

- Process:
  - Upload a sequence *fasta* file.
  - Select “similar” reference geno from dropdown.
  - Choose various optional proce (e.g. pseudochromosome conti with ABACAS).
  - Submit with (optional) email ad for notification.
- Outputs available for up to months online.
- Download outputs in EBI / GenBank compatible formats.



Step 5: Pseudochromosome contiguation

The contiguation step will try to orientate the sequences in your input file to align with the chromosomal sequences of the reference organism to build pseudochromosomes, which

Taxon ID 5653

Example: *Companion*

Database ID Companion

Your contact information (optional)

You can leave your email address if you want to be notified when your job starts and finishes. This is absolutely optional, if you choose not to share your email address, you can always manually check the status of your job using a private link provided by us after submission.

Email

To protect the service from automated bots, please prove that you are a human.

Enter the code displayed above.

gene models, but is not guaranteed to work for annotating genomes not closely related to the chosen reference.

● Yes, use RATT with the Species transfer type to transfer reference gene models.

○ No, only do *ab initio* gene finding.

Companion was first developed in 2015 by the Parasite Genomics group at the Wellcome Sanger Institute.

Ongoing development has been carried out by the biottofni group at the University of Glasgow since 2017.

# NEW FEATURES

CURRENTLY IN DEVELOPMENT

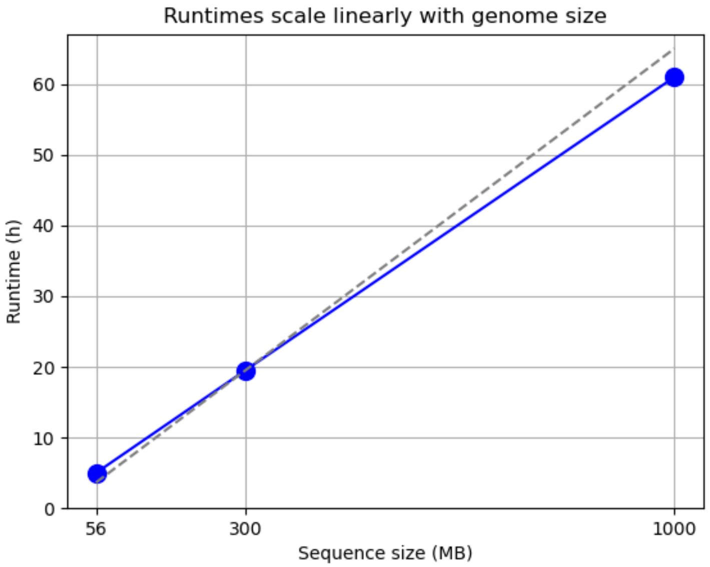
# SCALING FOR LARGER GENOMES

## References gathered from VEuPathDB projects and pre-compiled.

- In production: protozoa, fungi.
- In development: **VectorBase**
- Exploring: **HostDB**

VectorBase
Bioinformatics Resource for
Invertebrate Vectors of Human Pathogens

HostDB
Release 57
21 Apr 2022
Pathogen Host Informatics Resources

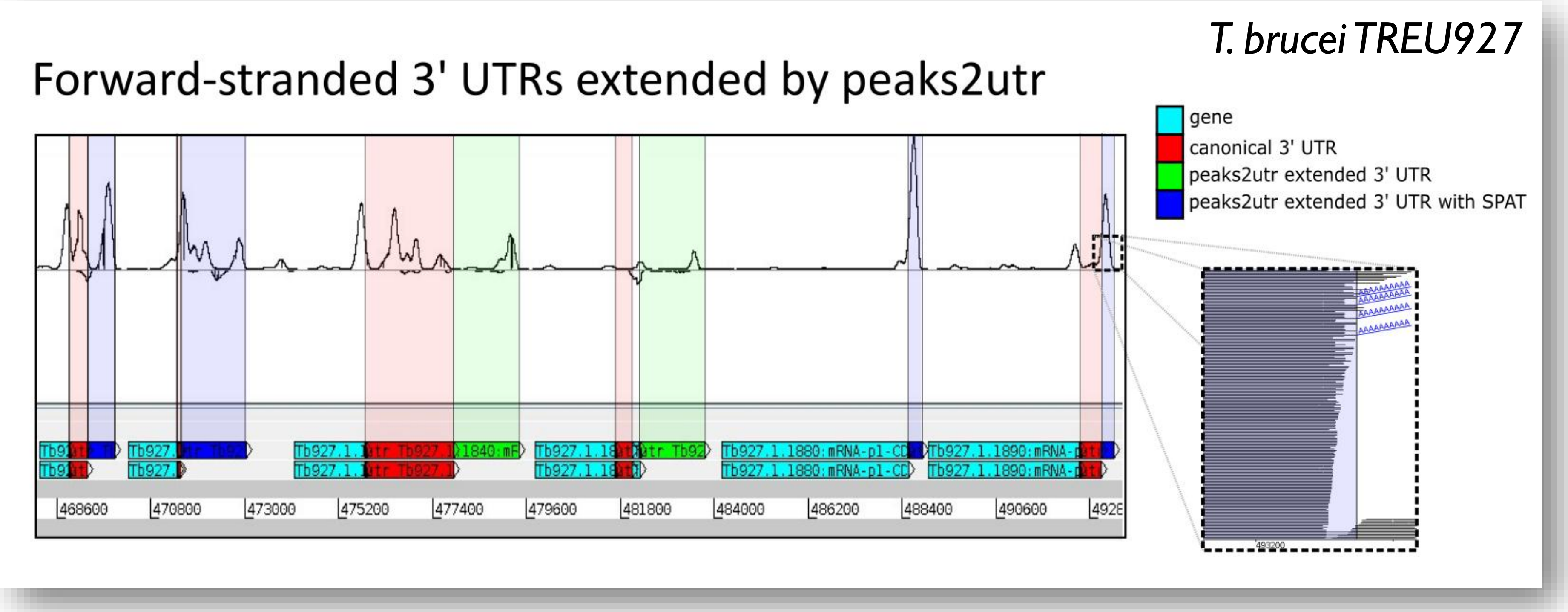


- Tested several VectorBase reference genomes
- Runtimes scale linearly
- Issues with RATT for larger genomes
  - Explore fix or alternative tool (such as *Liftoff*)

Alaina Shumate, Steven L Salzberg, **Liftoff: accurate mapping of gene annotations**, *Bioinformatics*, Volume 37, Issue 12, 15 June 2021, Pages 1639–1643, https://doi.org/10.1093/bioinformatics/btaa1016

# UTR ANNOTATION

- **peaks2utr**: stand-alone 3’ UTR annotation Python tool
- How to incorporate:
  - Apply to select reference annotations - or to output annotations?



*T. brucei TREU927*

Forward-stranded 3' UTRs extended by peaks2utr

# FUTURE ADDITIONS

- Improved efficiency / job concurrency.
- Expanded reference set.
- Dynamic reference updates.
- Improved visualisations (e.g. Apollo).
- Full project submission for EBI / GenBank
- Additional functionality to increase richness of annotations.

# CELL ATLAS – cell×gene

## RESPONSIVE WEB INTERFACE FOR VISUALISING SINGLE CELL DATA

cell×gene

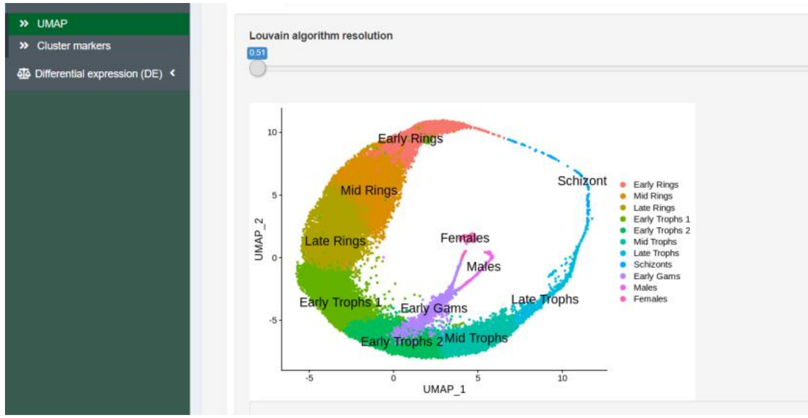
# WHAT IS CELLXGENE?

- Interactive single cell (SC) analysis framework.
- Developed by Chan Zuckerberg Initiative for the Human Cell Atlas.
- Predominantly Python with Flask app interface.
- Simple to install and launch from command line.
- Third-party tools for hosting (cellxgene-gateway) and enhancing featureset (cellxgene_VIP, excellxgene).

# COMPARISON

## UofG III - Cell Atlas

- R/Shiny interface.
- Must be maintained by in-house developers at UofG.
- Each dataset requires a bespoke approach.
- Manually defined plotting groups.

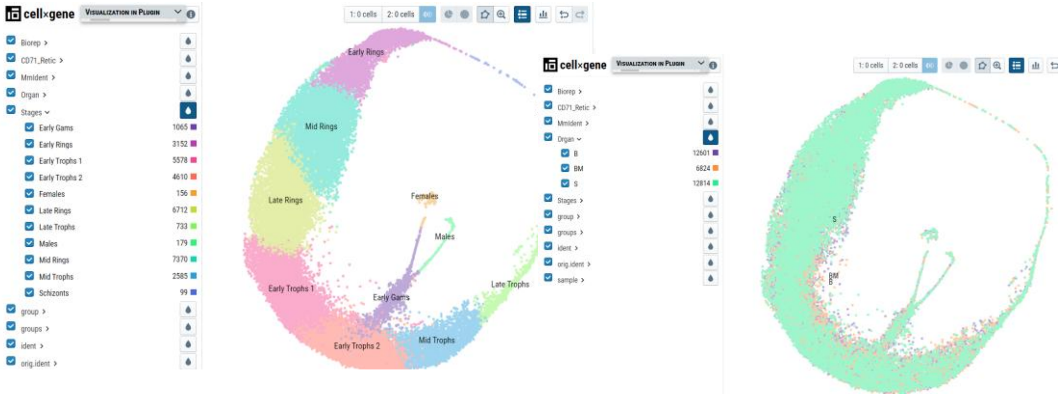


Louvain algorithm resolution: 0.51

- Unresponsive interface effects user experience.

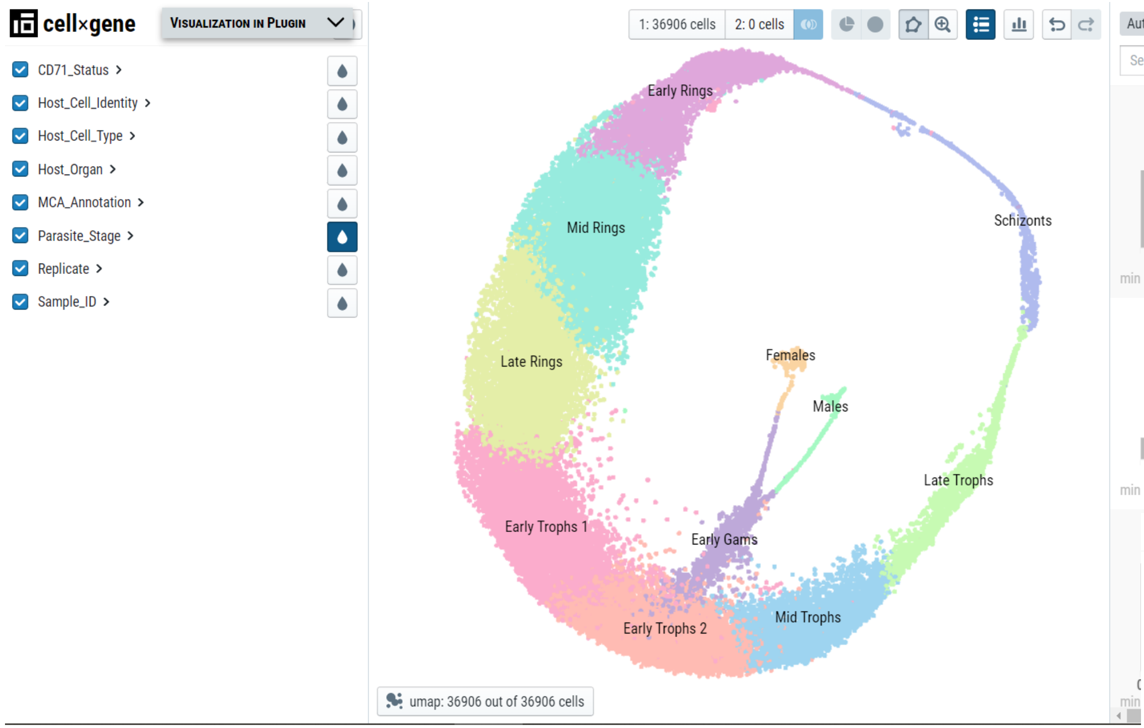
## Cell Atlas - cellxgene

- + Python/Flask interface.
- + Regularly maintained GitHub repos.
- + All functionality included out-of-the-box for each dataset.
- + Plotting groups dynamic and inferred from dataset.



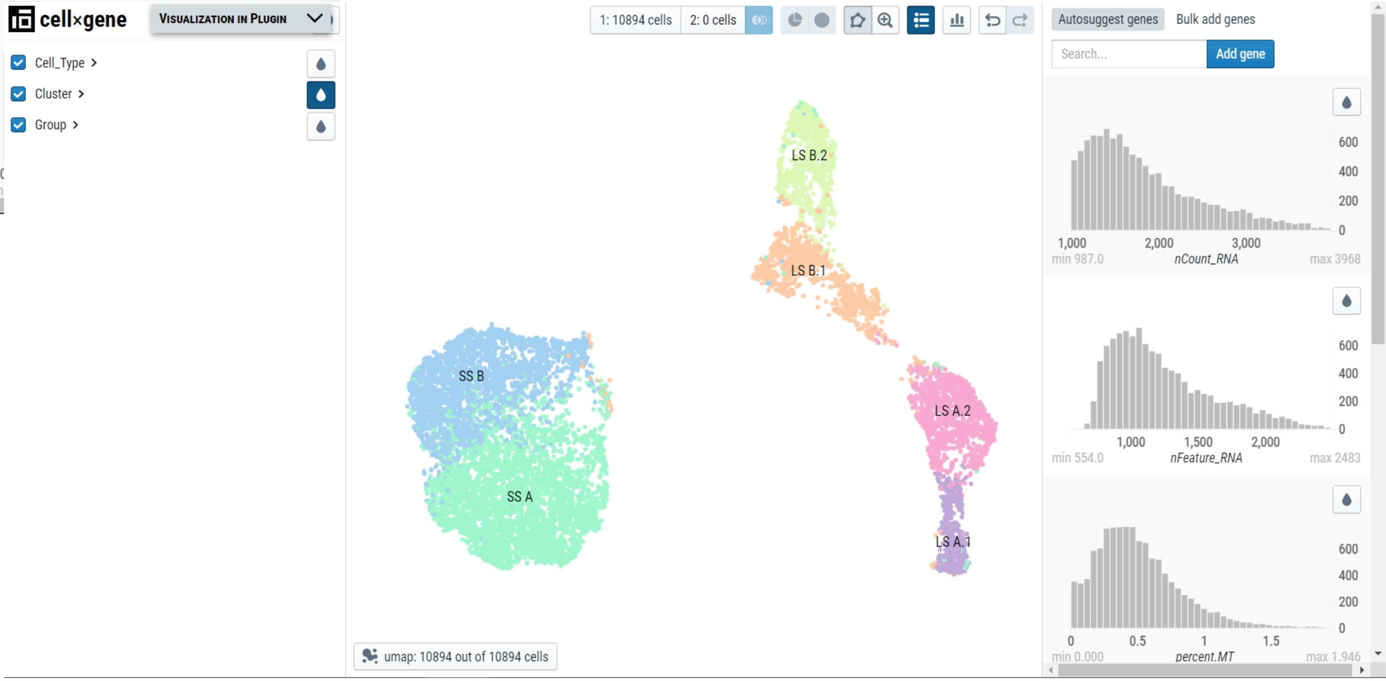
- + Once dataset loaded, interface is snappy.
- + Includes all the same features, and then some.

cell×gene



https://cellatlas-cxg.mvls.gla.ac.uk/Tbrucei/

*Trypanosoma brucei*



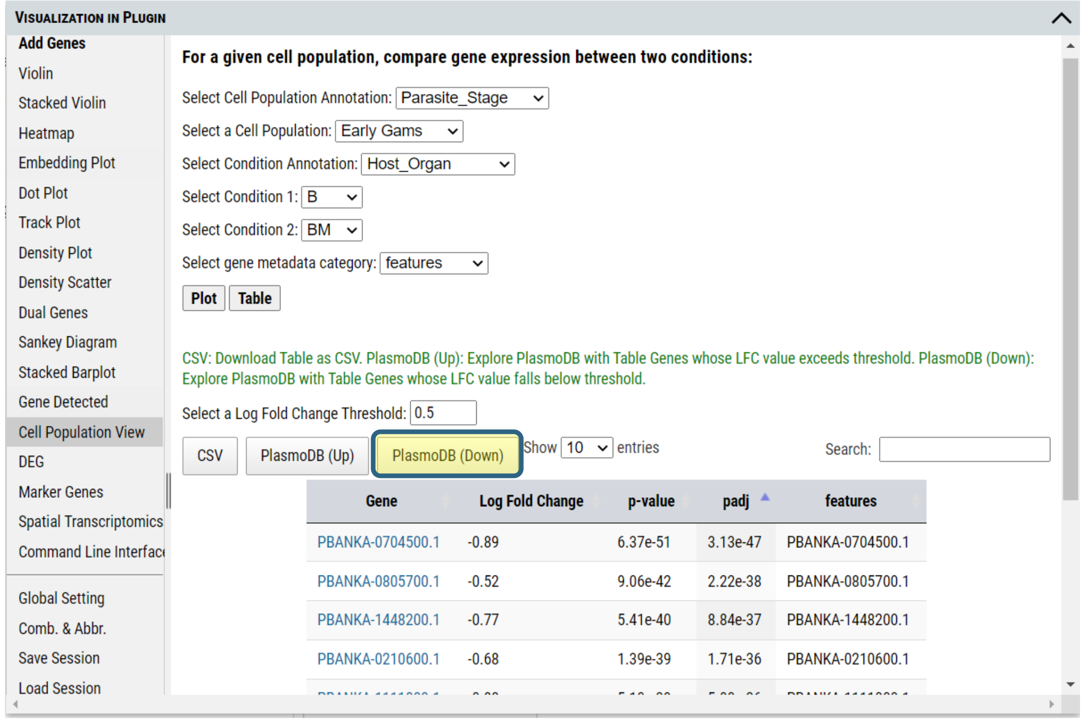
*Plasmodium berghei*

https://cellatlas-cxg.mvls.gla.ac.uk/Pb/

# ADDITIONAL FEATURES

CURRENTLY IN DEVELOPMENT

# VEUPATHDB INTEGRATION



**For a given cell population, compare gene expression between two conditions:**

Select Cell Population Annotation: Parasite_Stage

Select a Cell Population: Early Gams

Select Condition Annotation: Host_Organ

Select Condition 1: B

Select Condition 2: BM

Select gene metadata category: features

Plot | Table

CSV: Download Table as CSV. PlasmoDB (Up): Explore PlasmoDB with Table Genes whose LFC value exceeds threshold. PlasmoDB (Down): Explore PlasmoDB with Table Genes whose LFC value falls below threshold.

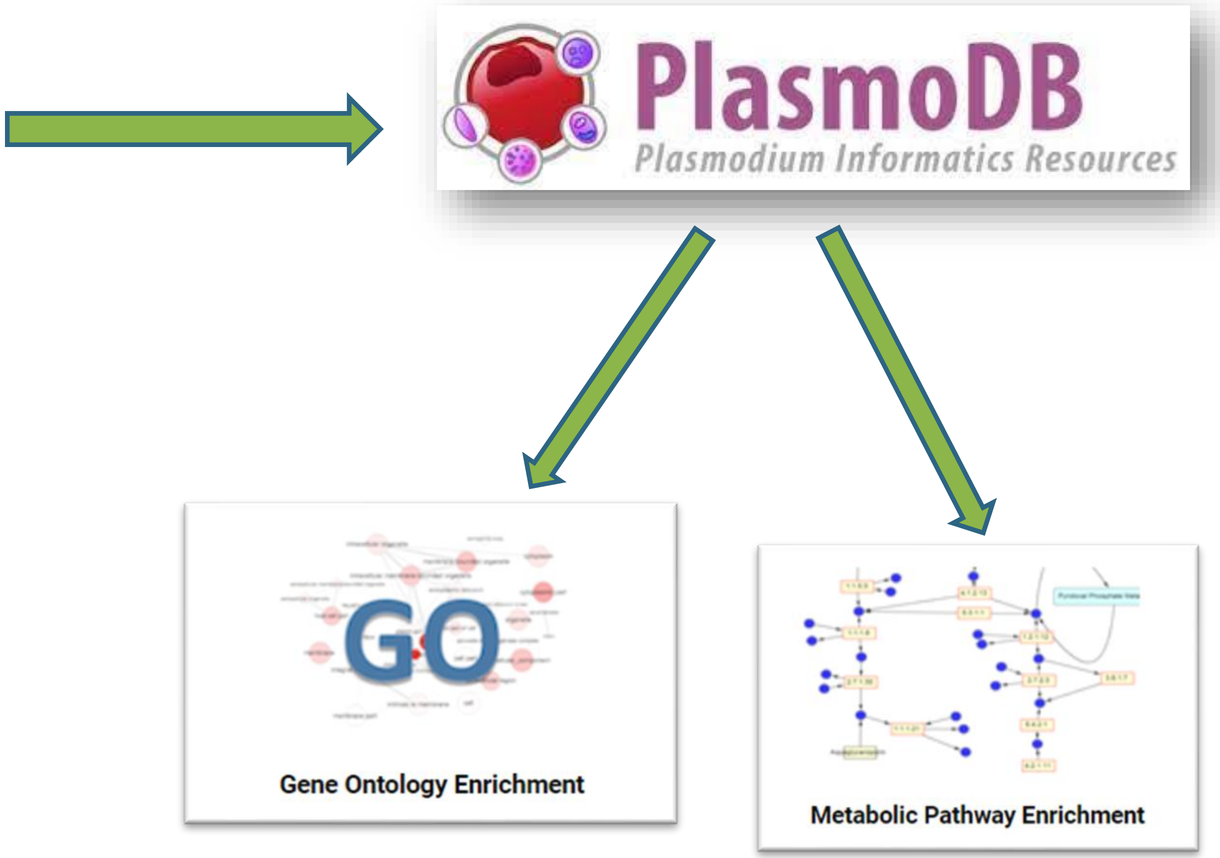
Select a Log Fold Change Threshold: 0.5

CSV | PlasmoDB (Up) | PlasmoDB (Down)

Show 10 entries

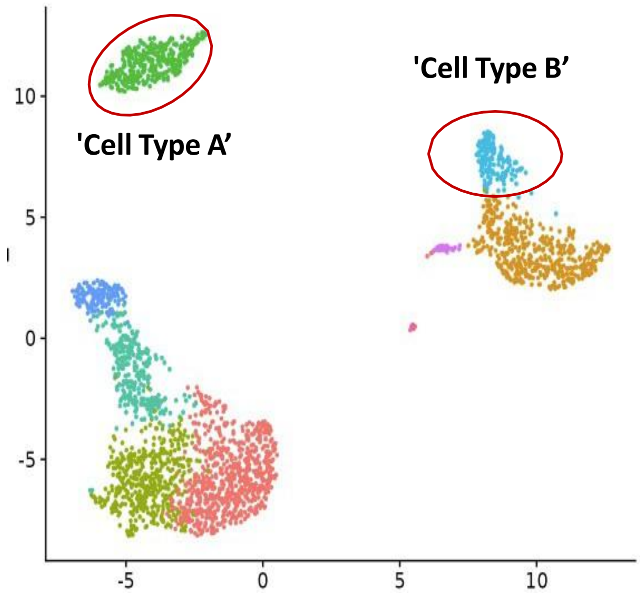
Search:

| Gene | Log Fold Change | p-value | padj | features |
|---|---|---|---|---|
| PBANKA-0704500.1 | -0.89 | 6.37e-51 | 3.13e-47 | PBANKA-0704500.1 |
| PBANKA-0805700.1 | -0.52 | 9.06e-42 | 2.22e-38 | PBANKA-0805700.1 |
| PBANKA-1448200.1 | -0.77 | 5.41e-40 | 8.84e-37 | PBANKA-1448200.1 |
| PBANKA-0210600.1 | -0.68 | 1.39e-39 | 1.71e-36 | PBANKA-0210600.1 |

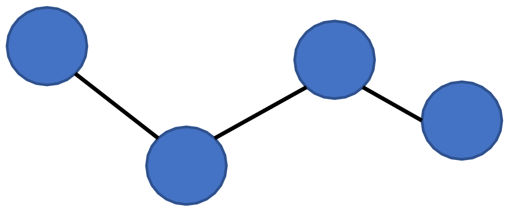

PlasmoDB
Plasmodium Informatics Resources
GO
Gene Ontology Enrichment
Metabolic Pathway Enrichment

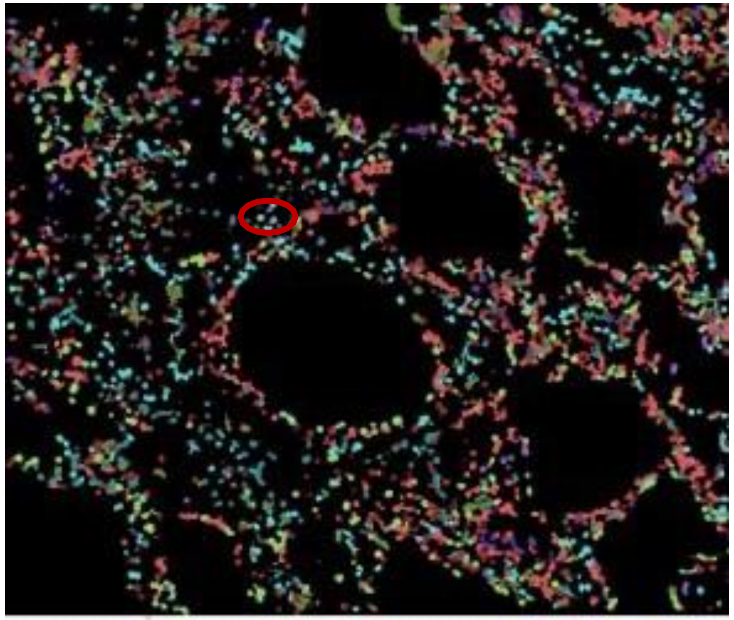
- Gene detection algorithm
  - – link to various VEuPathDB sub-sites

# EXPLORATION AND VISUALIZATION OF CELL-CELL INTERACTIONS*



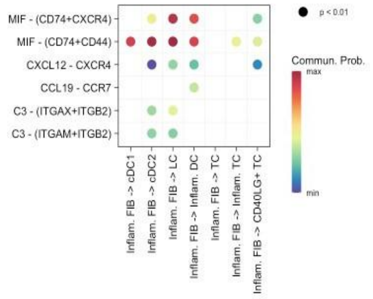
Combined projection of interacting cells

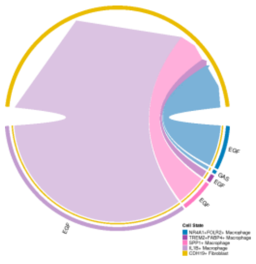




Single cell-RNA sequencing data

Spatial data





User customised plots

* PhD project of Olympia Hardy

RESEARCH INTO INFLAMMATORY ARTHRITIS CENTRE
VERSUS ARTHRITIS

# ACKNOWLEDGEMENTS

Application Deadline: 07/07/2022

Thomas Otto
Kathryn Crouch
Edward Agboraw
Olympia Hardy
Scott Arkison

Ross Laidlaw
Fiona Achcar
Alex Pancheva
John Cole

Lucy MacDonald
Eva Crespo
Lauren Galloway
Collins Morang'a

Emma Briggs
Domenico Somma
Theodore Simakou
Katie Chapple