



University
of Glasgow

Laidlaw, Ross (2024) *Applications of scRNA-seq to capture Trypanosomatid life cycle stage transitions and heterogeneity*. PhD thesis.

<https://theses.gla.ac.uk/84743/>

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses

<https://theses.gla.ac.uk/>
research-enlighten@glasgow.ac.uk

Applications of scRNA-seq to capture Trypanosomatid life cycle
stage transitions and heterogeneity

Ross Laidlaw
BSc (Hons)

September 25, 2024

Submitted in fulfilment of the requirements for the
Degree of Philosophy

Institute of Infection and Immunity,
College of Medical and Veterinary Sciences

University of Glasgow

1 Abstract

Single cell RNA-sequencing has transformed transcriptomic analysis, giving researchers unparalleled access to minute changes in gene expression at the level of single cells. This resolution is especially useful when analysing biological processes, as cells (and their transcriptome) from various points in the process can be captured. The applications of scRNA-seq to parasite life cycles is thus clear. In this thesis, I have applied scRNA-seq and a variety of analysis methods to interrogate the transitions that occur in the Trypanosomatid group of parasites, specifically *Trypanosoma brucei* and *Trypanosoma cruzi*. Firstly, a tool was created that aligns and compares biological processes across time between two conditions. This tool returns more accurate alignments of asymmetrical processes than current tools and allows extraction of meaningful genes that define the differences between the two conditions. Next, we confirmed the effectiveness of scRNA-seq to profile the transcriptome of *T. cruzi*, in the process capturing marker genes of the four major life cycle stages of the parasite and creating the first scRNA-seq atlas of *T. cruzi*. From this data the process of metacyclogenesis was captured at the single cell level, characterising various differentially expressed genes by their expression peak across the process and identifying several differentially expressed RNA binding proteins which had not previously been associated with metacyclogenesis. The transcriptomic profile of trypomastigote subsets was investigated, possibly identifying functional trypomastigote subsets in the context of infection. Finally, we utilised scRNA-seq on the *RBP6* overexpression *in vitro* model of the insect stage life cycle of *T. brucei* to identify genes which drive the transition of life cycle stages. These results revealed an unexpected lack of transcriptomic signature epimastigotes being generated by the model but did capture the generation of few metacyclic-like cells who express mVSG but have many differences in transcriptome when compared to the *in vivo* forms, including metabolism associated genes. PCF-like cells, high in expression of PAG, were also detected in the data, with the transcriptome of these cells being highly expressed towards the end of the *RBP6* overexpression process in published bulk RNA-seq data of the model.

Contents

1	Abstract	2
2	Acknowledgements	7
3	Author's Declaration	8
4	Abbreviations	9
5	Introduction	12
5.1	ScRNA-sequencing technologies	12
5.2	Processing pipeline of Single Cell RNA-sequencing	13
5.3	Integrating scRNA-seq datasets	18
5.4	Trajectory inference	20
5.5	ScRNA-seq analysis of parasites	22
5.6	Introduction to Kinetoplastids	22
5.7	The life cycle of <i>T. brucei</i>	23
5.8	The life cycle of <i>T. cruzi</i>	26
5.9	Problems with scRNA-seq analysis of Trypanosomatids	29
6	Aims and Objectives	30
7	Chapter 1: TrAGEDy - Trajectory Alignment of Gene Expression Dynamics	30
8	Chapter 1: Introduction	30
9	Chapter 1: Results	31
9.1	Aligning simulated scRNA-seq datasets	34
9.2	WT vs <i>ZC3H20</i> KO <i>T. brucei</i> alignment	36
9.3	WT vs <i>Bcl11b</i> KO T cell development analysis	39
10	Chapter 1: Materials and Methods	42
10.1	The TrAGEDy method	42
10.1.1	Create interpolated points	42
10.1.2	Scoring dissimilarity	42
10.1.3	Identifying the optimal path of the two trajectories	42
10.1.4	Align pseudotime of interpolated points and cells	43
10.1.5	Differential expression analysis with TrAGEDy	44
10.2	Simulated dataset creation and benchmarking	44
10.2.1	Simulating trajectories with Dyngen	44
10.2.2	Applying TrAGEDy, Genes2Genes and cellAlign to simulated trajectories	45
10.3	Analysis of WT vs <i>ZC3H20</i> KO <i>T. brucei</i> scRNA-seq dataset	45
10.3.1	Pre-processing of WT vs <i>ZC3H20</i> KO <i>T. brucei</i> dataset	45
10.3.2	TrAGEDy analysis of WT vs <i>ZC3H20</i> KO <i>T. brucei</i> dataset	45
10.3.3	Seurat analysis of WT vs <i>ZC3H20</i> KO <i>T. brucei</i> dataset	45
10.3.4	TradeSeq analysis of WT vs <i>ZC3H20</i> KO <i>T. brucei</i> dataset	46
10.4	GO term analysis of <i>T. brucei</i> DE genes	46
10.5	Analysis of WT vs <i>Bcl11b</i> KO <i>in vitro</i> T cell development scRNA-seq dataset	46

10.5.1	Preprocessing of WT vs <i>Bcl11b</i> KO <i>in vitro</i> T cell development dataset prior to TrAGEDy analysis	46
10.5.2	TrAGEDy analysis of WT vs <i>Bcl11b</i> KO <i>in vitro</i> T cell development dataset	47
10.5.3	Seurat analysis of WT vs <i>Bcl11b</i> KO <i>in vitro</i> T cell development dataset	47
10.5.4	TradeSeq analysis of WT vs <i>Bcl11b</i> KO <i>in vitro</i> T cell development dataset	47
10.5.5	GO term analysis of T cell DE genes	48
10.6	Runtime experiments	48
10.7	Package versions	48
11	Chapter 1: Discussion	48
12	Chapter 2: <i>In vitro</i> single cell transcriptomic atlas of <i>T. cruzi</i> development	51
13	Chapter 2: Introduction	52
14	Chapter 2: Results	53
14.1	Data generation	53
14.2	Mapping and quality control of <i>T. cruzi</i> scRNA-seq data	53
14.3	Integrating the atlas samples	55
14.4	Bulk RNA-sequencing analysis of the <i>T. cruzi in vitro</i> life cycle	58
14.5	The <i>T. cruzi in vitro</i> atlas as a tool for annotating datasets	62
14.6	Cell cycle analysis of the <i>T. cruzi</i> atlas	64
14.7	Characterising trypomastigote heterogeneity	67
14.8	Identifying markers of <i>T. cruzi</i> metacyclogenesis	67
15	Chapter 2: Materials and Methods	71
15.1	Sample collection	71
15.1.1	Parasite culture	71
15.1.2	Sample collection	72
15.1.3	Cell viability assessment	72
15.1.4	Bulk RNA-seq – RNA extraction, library preparation and sequencing	73
15.1.5	Single-cell RNA-seq – library preparation and sequencing	73
15.2	Bioinformatics analysis	74
15.2.1	Creating reference transcriptomes for mapping	74
15.2.2	Mapping and counting of life cycle bulk RNA-seq samples	74
15.2.3	Processing and analysis of bulk RNA-seq samples	74
15.2.4	Mapping and counting scRNA-seq reads	74
15.2.5	Quality control and processing of scRNA-seq samples	74
15.2.6	Integration of <i>T. cruzi</i> atlas samples	75
15.2.7	Cell cycle analysis of the <i>T. cruzi</i> atlas	75
15.2.8	Analysis of epimastigote to metacyclic trypomastigote transitions in scRNA-seq data	76
15.2.9	Analysis of trypomastigote subsets	76
15.2.10	Reference mapping using the <i>T. cruzi</i> atlas	76
15.2.11	Annotating an unrelated scRNA-seq run using the <i>T. cruzi</i> atlas	76
15.2.12	Package versions	77
16	Chapter 2: Discussion	77
17	Chapter 3: Transcriptomic analysis of <i>RBP6</i> overexpression induced <i>T. brucei</i> insect life cycle stages	80

18 Chapter 3: Introduction	81
19 Chapter 3: Results	83
19.1 Implementing <i>RBP6</i> overexpressing PCFs	83
19.2 Image analysis of the Pasteur <i>RBP6</i> overexpressing cells	87
19.3 Single cell RNA sequencing data analysis of the <i>RBP6</i> overexpression model of <i>T. brucei</i> development	89
19.3.1 Capturing the transcriptome of <i>RBP6</i> overexpressing single cells	89
19.3.2 Overview and quality control of the <i>RBP6</i> overexpression scRNA-seq dataset	90
19.3.3 Regressed analysis of the <i>RBP6</i> overexpression scRNA-seq dataset	92
19.4 Integration of <i>RBP6</i> overexpression cells and <i>in vivo</i> insect life cycle stage cells	99
19.5 Integration of <i>RBP6</i> overexpression cells and <i>in vitro</i> culture PCFs	106
19.6 Integration consistency of <i>T. brucei</i> scRNA-seq datasets	111
20 Chapter 3: Materials & Methods	113
20.1 Preparation of PCF <i>T. brucei</i> culture media	113
20.1.1 Procedure to make Semi defined media 79 (SDM-79)	113
20.1.2 Procedure to make Semi defined media 80 (SDM-80)	113
20.2 <i>In vitro</i> culturing of PCF <i>T. brucei</i>	115
20.2.1 Calculating <i>T. brucei</i> culture concentrations	115
20.3 Gel electrophoresis	115
20.4 Gel extraction	116
20.5 Creating ampicillin-resistance agar plates	116
20.6 Assessing concentration of DNA	116
20.7 Creating the <i>RBP6</i> overexpression plasmid	116
20.7.1 Amplifying <i>RBP6</i> gene from <i>T. brucei</i> genomic DNA with polymerase chain reaction	116
20.7.2 Extracting pLEW100v5 plasmid backbone	117
20.7.3 Ligation of <i>RBP6</i> gene sequence into pLEW100v5 plasmid	117
20.7.4 Transformation of engineered <i>RBP6</i> overexpression plasmid into DH5- α <i>E. coli</i>	118
20.7.5 Extraction and purification of engineered <i>RBP6</i> overexpression plasmid from DH5- α <i>E. coli</i>	118
20.7.6 Sanger sequencing of <i>RBP6</i> overexpression plasmid insert	119
20.7.7 Transfection of engineered <i>RBP6</i> overexpression plasmid into 29:13 <i>T. brucei</i> cells	119
20.8 Inducing overexpression of <i>RBP6</i> in 29:13 <i>T. brucei</i> cells	120
20.9 Immunofluorescence assay of <i>RBP6</i> overexpressing <i>T. brucei</i> PCF cells	120
20.10 Imaging of <i>RBP6</i> overexpression IFA slides	121
20.11 Single cell RNA sequencing of <i>RBP6</i> overexpressing 29:13 <i>T. brucei</i> cells	121
20.12 Mapping the <i>RBP6</i> overexpression single cell RNA-seq data	122
20.13 Analysis of the <i>RBP6</i> overexpression dataset	122
20.13.1 Quality control and preprocessing	122
20.13.2 Cell cycle labelling of the <i>RBP6</i> overexpression scRNA-seq dataset	123
20.13.3 Analysis of the <i>RBP6</i> overexpression process	123
20.13.4 Comparison of scRNA-seq and bulk RNA-seq <i>RBP6</i> overexpression data	124
20.13.5 Cluster annotation of <i>RBP6</i> overexpression scRNA-seq data	124
20.14 Integration of <i>RBP6</i> overexpression scRNA-seq data and <i>in vivo</i> insect stage scRNA-seq data	125
20.15 Integration of <i>RBP6</i> overexpression scRNA-seq data and <i>in vitro</i> PCFs	125
20.16 Testing consistency of integration methods on <i>T. brucei</i> scRNA-seq data	126
20.17 Package Versions	126

21 Chapter 3: Discussion	126
22 Conclusion and future directions	133
23 Code Availability	136
24 Appendices	136
24.1 Appendix files	136
24.2 Appendix tables	136
24.3 Appendix figures	137
25 References	154

2 Acknowledgements

The first thanks goes to my supervisors. Thomas Otto for not letting me give up on anything (no matter how much I wanted to) and for facilitating and encouraging my bioinformatics journey over the past 5 years. I hate to say it, but you were probably right to not give up on some of it. Probably. Richard McCulloch for always encouraging me in whatever I choose to pursue while making sure I still had my eyes firmly set on the goals of the thesis. Keith Matthews for making sure that I always had an eye open for the small details and for asking questions which cut through to the heart of the hypothesis. Last, but not least, Emma Briggs for laying the groundwork of scRNA-seq analysis of Trypanosomatids so well (I would have been lost without it) and providing so much support in every aspect of my PhD.

I'd like to thank the Otto lab (and Otto lab adjacent) members (Olympia, Andrew, Yiyi, Brenda, Joy, Will, Fiona, Dom, Jack, Lucy, Toby and Sam) for creating such a positive and friendly lab environment; I'm sad to be leaving it. Special thanks to Olympia, Andrew, Brenda, Joy, Fiona and Yiyi for making our office the best office in the building.

I'd like to especially thank the first Otto lab PhD student: Alexandrina Pancheva, who showed me the ropes of bioinformatics and put up with my constant questions and complaints. Its fair to say I wouldn't be where I was without your guidance, although I would perhaps be a less pessimistic person.

This PhD project would not be possible without the constant IT support of Scott Arkison, a man who single handedly makes all scientific research possible in MVLS.

Many thanks to Lucy Glover for allowing me to come visit and work in her lab. It was one of the stand out moments of my PhD and my biggest regret that the project didn't go anywhere.

I'd like to thank everyone in the McCulloch lab, past and present (Craig, Marija, Grace, Jeziel, Jane, Catarina, Mark and Gabriel). I would like to especially thank Catarina Almeida de Marques, Jane Munday and Mark Girasol for providing their much needed (and appreciated) support in the wetlab and Gabriel Lamak Almedia da Silva for calmly and clearly explaining how to use the Leica machine and Fiji.

I'd also like to thank the Matthews lab, past and present (Kirsty, Mathieu, Frank, Steve, Guy, Ruth, KK, Balazs, Federico) for putting up with my rambling bioinformatics presentations and for hosting some of the best lab nights out possible.

I'd like to give thanks to my parents, who have never erred to support me in pursuing this project, or anything else in life. I love you both.

To my partner Charlotte, not only I but everyone who reads this thesis owes thanks to you for being the person who managed to copywrite my manic ramblings into something of actual substance and coherence. Thank you for being there for me through this process and supporting me every step of the way.

Finally, I would like to dedicate this thesis to my Grandma, without whom I would not be myself in any way.

3 Author's Declaration

I here declare that this thesis and the results presented in it, are the result of my own work, except where otherwise stated and acknowledged. None of the results presented have been previously used to obtain a degree at any university.

Ross France Laidlaw, August 2024

4 Abbreviations

- ADT: amastigote-derived trypomastigotes sample
- AMA: amastigote sample
- BARP: Brucei stage Alanine-rich Protein
- BBKNN: Batch Balanced N Nearest Neighbours
- BSF: Bloodstream Form
- cAMP: 3',5'-cyclic adenosine monophosphate
- CCA: citrate/cis aconitate
- CRD: Cross-reacting determinant
- DE: Differentially Expressed
- DNA: Deoxyribonucleic acid
- DN: Double Negative
- DP5CDH: delta-1-pyrroline-5-carboxylate dehydrogenase
- DTU: Discrete Type Unit
- DTW: Dynamic Time Warping
- ESAG: Expression site-associated gene
- *E. coli*: *Escherichia coli*
- EP: epimastigote sample
- FA2: ForceAtlas2
- fMNIST: fashion Modified National Institute of Standards and Technology database
- GAM: General Additive Model
- gDNA: Genomic DNA
- G1e: G1 early
- G2G: Genes2Genes
- G1l: G1 late
- G2M: G2/Mitosis
- GO: Gene Ontology
- GP63: Glycoprotein 63
- GP82: Glycoprotein 82

- HF: High Fidelity
- hpi: hours post infection
- IFA: Immunofluorescence Assay
- ISG: Invariant Surface Glycoprotein
- KBL: 2-amino-3-ketobutyrate coenzyme A ligase
- kDNA: kinetoplastid DNA
- kNN: k-Nearest Neighbours
- KO: Knockout
- LDH: L-threonine 3-dehydrogenase
- Log₂FC: Log₂ Fold Change
- LS: Long Slender
- MASP: Mucin-Associated Surface Protein
- mVSG: Metacyclic VSG
- NK: Natural Killer
- OE-plasmid: *RBP6* overexpression plasmid
- ORF: Open Reading Frame
- PAG: Procyclin Associated Gene
- PAD: Protein Associated with Differentiation
- PHATE: Potential of Heat diffusion for Affinity-based Transition Embedding
- PCA: Principal Component Analysis
- PC: Principal Component
- PCR: Polymerase chain reaction
- PCF: Procyclic Form
- PDH: Proline dehydrogenase
- PM: Peritrophic Matrix
- PV: Parasitophorous Vacuole
- RBP: RNA binding protein
- RDS: R Dataset
- RHS: Retrotransposon Hot Spot

- RNA: Ribonucleic acid
- rRNA: ribosomal RNA
- rpm: revolutions per minute
- Seurat V5 CCA: Seurat V5 Canonical Correlation Analysis
- SIA: Sialic Acid
- scRNA-seq: single cell ribonucleic acid sequencing
- SS: Short Stumpy
- SE/MT: stationary phase epimastigote/metacyclic trypomastigote sample
- TAE: Tris-acetate-Ethylenediaminetetraacetic acid
- TCR: T Cell Receptor
- TDH: L-threonine 3-dehydrogenase
- TI: Trajectory Inference
- TrAGEDy: Trajectory Alignment of Gene Expression Dynamics
- *T. brucei*: *Trypanosoma brucei*
- *T. cruzi*: *Trypanosoma cruzi*
- tSNE: t-distributed Stochastic Neighbor Embedding
- UMAP: Uniform Manifold Approximation and Projection
- UMI: Unique Molecular Identifier
- VSG: Variable Surface Glycoprotein
- WT: Wild type
- ZC3H: Zinc Finger Protein

5 Introduction

Even within the same tissue or life cycle stage, cells exhibit heterogeneity in terms of gene expression (H. Chen et al. 2019, Briggs, Marques, et al. 2023, Smircich, Perez-Diaz, et al. 2023). This heterogeneity makes it unwise to analyse the transcriptome of tissues and stages at a bulk level, as this transcriptomic heterogeneity is lost when the mRNA contents of all the cells are combined together into an additive sample. To make sure this heterogeneity is preserved, cells will need to be assessed as individuals, rather than as a bulk.

5.1 ScRNA-sequencing technologies

As early as 2002, methods have existed to capture the transcriptome of single cells (Levsky et al. 2002). At the beginning of single cell RNA-sequencing’s (scRNA-seq) life, however, the number of genes that could be captured was incredibly low, nowhere near the level of whole transcriptome, which could be achieved with bulk RNA-seq; meaning scRNA-seq had some way to go to truly realise its potential. It was not until 2009 when the first whole transcriptome scRNA-seq analysis was carried out by Tang and colleagues (Tang et al. 2009) on mouse embryos. To capture the whole transcriptome of a single cell, the authors first extracted a single cell from a mouse embryo and lysed it to release the contents of the cell. Poly adenylated sequences were then converted into cDNA through reverse transcription and sequenced, before mapping and counting of the amplified fragments was carried out to quantify the amount of mRNA within the cell. There are of course downsides to Tang and colleagues’ methods including low throughput and PCR amplification bias possibly causing noise in gene expression profiles (S. Li et al. 2014, Dijk, Jaszczyszyn, and Thermes 2014, H. Shi, Y. Zhou, et al. 2021). Fortunately, modern scRNA-seq methods have ways to mitigate these issues. For the issue of PCR amplification one of the most common methods is the addition of a Unique Molecular Identifier (UMI) to the mRNA molecules prior to sequencing; UMIs can be used to identify which amplified sequence originates from the same mRNA by matching their UMI and collapsing them into one when it comes to calculating gene expression counts (Kivioja et al. 2011). Interestingly, some papers suggest that the effect of amplification bias (due to lack of UMI) on differential expression is only mild to negligible (Parekh et al. 2016). For low throughput, droplet-based methods (such as 10X Chromium and drop-seq (Macosko et al. 2015)) are applied to separate out the single cells using microfluidics. Briefly, cells are sent through a microfluidic system where they meet with a bead coated in oligonucleotides, before both are encapsulated in oil. Within this oil droplet, the cells are lysed and the mRNA is captured through hybridisation with the oligonucleotides. This makes the throughput of droplet-based methods high, but the need to rely on microfluidics systems means there is limited capacity to customise the process

An alternative to droplet-based methods are plate-based methods, such as SMART-seq3 (Hagemann-Jensen, Ziegenhain, P. Chen, et al. 2020) and MARS-seq (Keren-Shaul et al. 2019), on the other hand, use Fluorescence-activated Cell Sorting to separate individual cells into wells on a plate. The use of a plate places a limit on the number of cells that can be sequenced compared to droplet-based methods and the protocols have lower throughput than droplet-based methods. Plate-based method protocols however are more customisable, for example it is possible to deplete ribosomal RNA (rRNA) before sequencing occurs (Reid et al. 2018). Furthermore, transcript information is richer in plate-based methods (e.g. SNP and isoform information (Hagemann-Jensen, Ziegenhain, and Sandberg 2022)) as many of them allow the full length of the transcript to be captured (Probst et al. 2022).

All of the previous discussion is not to say that modern scRNA-seq techniques are without issues; there are many of them. One of the key issues with the technology is the prevalence of dropout events, where a gene is expressed in a cell but is not captured during sequencing (Haque et al. 2017). The

reasons for this range from the efficiency of mRNA capture to low expression of genes and, whatever the reasons, this leads to scRNA-seq data having a high number of zero values and uncertainty over whether a zero is biological (the gene is not expressed in the cell) or technical (a dropout event has occurred) (Jiang et al. 2022).

Going forward, the methods and issues discussed surrounding analysis of scRNA-seq data, or the data itself, will be in reference to 10X single cell RNA-sequencing (unless stated otherwise) as this is the only type of scRNA-seq data analysed in the course of this thesis.

5.2 Processing pipeline of Single Cell RNA-sequencing

The data generated using scRNA-seq is said to be high dimensional and sparse. High dimensional as it contains information on the expression of thousands of genes within thousands of cells, and sparse because it contains many zero values (P. Qiu 2020).

Thus, scRNA-seq data requires many preprocessing steps to allow downstream analysis to be completed accurately. As analysis of scRNA-seq is still a relatively new field, pipelines are still under development to best analyse the data, though there are certain steps which are essential to the analysis of scRNA-seq and some which are only essential depending on the context of the analysis. A suggested pipeline of scRNA-seq data analysis is outlined in figure 1, showing essential and optional processing of scRNA-seq data analysis. This figure will be commented on and referenced to throughout the rest of this section.

The initial steps of scRNA-seq analysis can be categorised as quality control measures. When cells are being processed, some may be in the process of dying; releasing their contents into the cell suspension solution. When the cells are encapsulated in their oil droplet, mRNA (known as ambient mRNA) released from dead/dying cells could also be captured in the bubble and can ultimately be amplified and captured in the final count information for that cell, contaminating the results (Fig. 2A). Methods like SoupX and DecontX (Yang et al. 2020, Young and Behjati 2020) aim to eliminate the contamination by adjusting the count matrices to remove ambient mRNA counts by utilising information contained in the raw count matrices, and by modelling the likelihood that counts come from cell-specific mRNA or ambient mRNA, respectively (Fig.1A). As these methods alter the count matrices directly, any incorrect adjustment of the matrices will be propagated through the rest of the analysis and into the downstream results, meaning caution should be taken when utilising these tools. The dead/dying cells themselves can still be captured in the droplets (Fig.1D) and droplets can exist with just ambient mRNA (Fig.1C). These types of low quality cell need to be removed from the data. To remove the latter from the data, filters are applied to remove cells with low numbers of UMI and gene counts (Fig.1B). The former type of low quality cell is removed through assessing the percentage of reads coming from the mitochondria, with the assumption that damaged/dying cells will have compromised cell membranes which cause cytoplasmic mRNA to leave the cell, while the mitochondrial mRNA remains (Fig. 2D, Illicic et al. 2016)

The final type of low quality cell are doublets, where two cells are encapsulate by the same oil droplet (Fig.2B). To remove doublets the UMI and gene counts are looked at, with the assumption that cells with high amounts are likely to be doublets (Fig.1B). Some doublets may be made up of cells with low mRNA counts, so only removing cells with high levels of mRNA does not mean all doublets will be caught. To try and better identify doublets in scRNA-seq data, a variety of doublet detection tools exist (Xi and J. J. Li 2021) (Fig.1C).

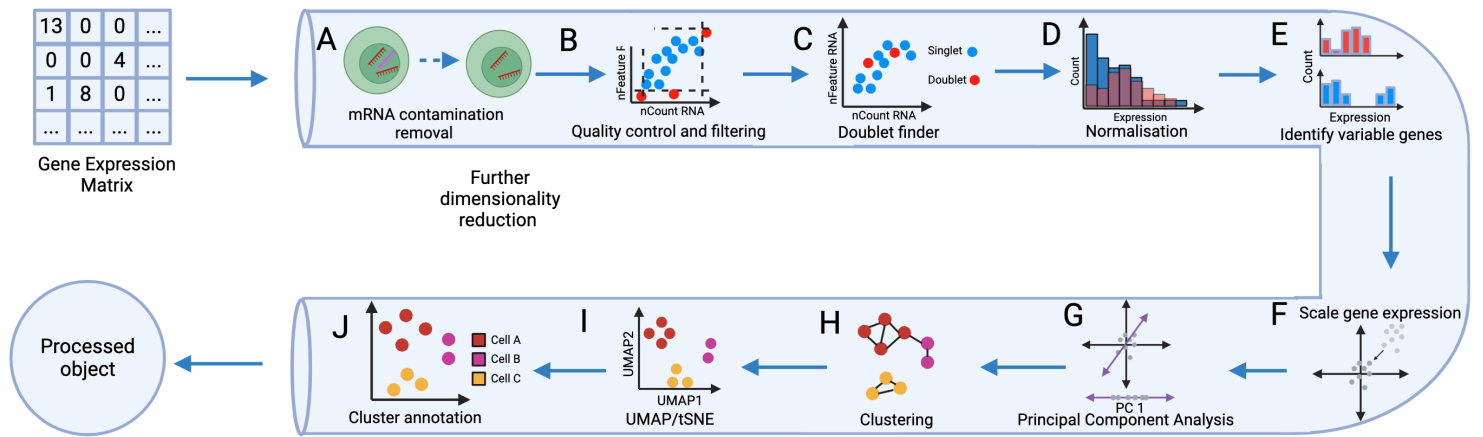


Figure 1: **A pipeline of scRNA-seq analysis**

Graphical representation of the various steps required to analyse a scRNA-seq dataset. A variety of quality control methods can be applied to the expression data, such as removing contaminating mRNA (A), dead/dying cells (B) and doublets (C). The raw gene expression values (blue) are then normalised (red) to make cells comparable and remove expression outliers (D). Genes that are highly (blue) and lowly (red) variable in terms of their expression are identified, with the top n highly variable genes being used for further downstream processing (E). The normalised expression values of the highly variable genes are then scaled (F) before Principal Component Analysis (PCA) is carried out on the scaled expression values (G). The PC space is then used as the basis for clustering (H) and further dimensionality reduction (I). Finally, the clusters are annotated by their cell type (or other such labelling strategy) (J)

There is a large degree of variability in terms of total mRNA content of a given cell (Marinov et al. 2014). This difference can be captured in the scRNA-seq data itself and can prove to be a bias when not addressed before downstream analysis is carried out. This is due to cells with higher mRNA counts having higher gene expression values and thus having a larger impact on the downstream processes, like differential expression. To deal with this effect, the raw gene expression data is normalised. The most commonly used normalisation method in scRNA-seq analysis is known as "library size" normalisation, where the read counts are normalised by a cell's library size, i.e. the total number of mRNA molecules present in a given cell, and is then scaled by a factor, either by a constant or the median of the total counts of all cells in the dataset, and log transformed (Fig.1D). This ensures that expression values can be expressed as a proportion of the scale factor. This type of normalisation method is naive, as it takes no sources of variation into account apart from total mRNA count, and thus could be influenced by many different forms of variation that are not related to the total mRNA count. Examples of such sources of variation could be those of sex, ethnicity, percentage of mitochondrial reads per cell, among others. As such, more complex normalisation methods have been devised such as scTransform, which normalises the data by modelling the gene expression of a given dataset and can take in (and adjust for) covariates, when modelling gene expression (Hafemeister and Satija 2019).

In the community, there has been much debate on which methods perform best for normalisation. Arguably, the two biggest recent benchmarks of normalisation methods come from Boeshaghi & colleagues (Boeshaghi et al. 2022) and Ahlmann-Eltze & Huber (Ahlmann-Eltze and Huber 2023), although they

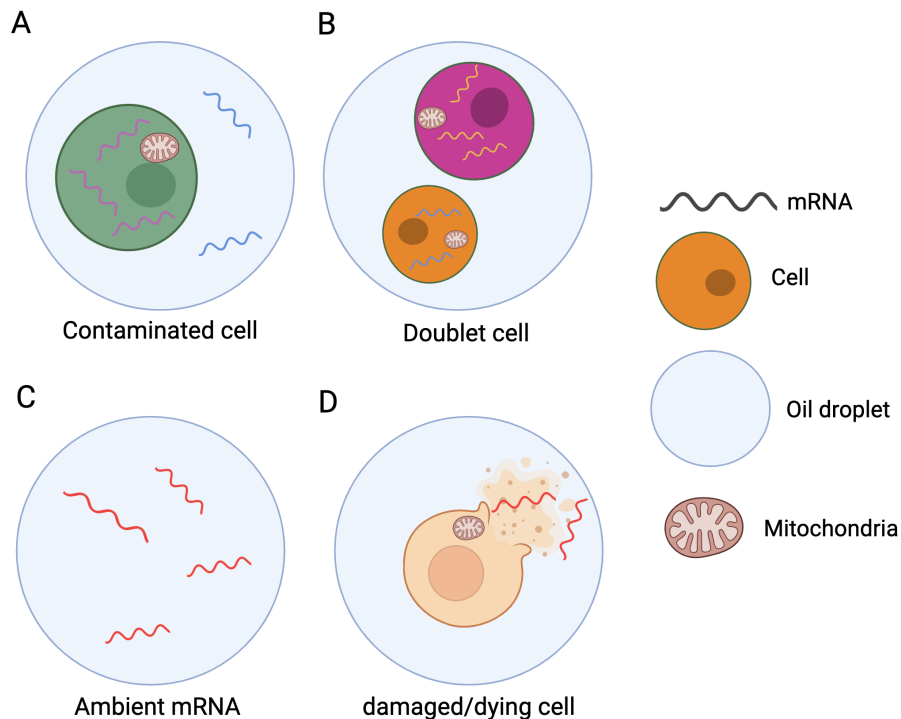


Figure 2: **Graphical example of low quality cells in scRNA-seq datasets**

Low quality cells can take on many forms in scRNA-seq data. Some cells could have mRNA contamination from other cells within its oil droplet (A), some could be two cells contained within the same droplet (doublets) (B), others could just be a droplet that contains only mRNA (C) and some cells may dead or damaged (D).

are by no means complimentary in terms of their findings. Boeshaghi & colleagues suggest that proportional fitting of the data, followed by log transformation and another round of proportional fitting, is the most optimal way normalise scRNA-seq data, while Ahlmann-Eltze & Huber disagree, concluding that it does not properly remove the effects of sequencing depth. Instead, they come to the conclusion that standard "library size" normalisation performs best. This inability to reach a consensus on a gold-standard normalisation method is further confounded by each benchmark paper designing their own metrics on how to best assess how well normalisation is performing.

The next step in the scRNA-seq data analysis pipeline is to identify highly variably expressed genes in the dataset, with these genes being the basis by which that the cells will be grouped and displayed (Fig.1E). The reasons for this are two-fold. First, by only using a subset of the genes in the dataset, it reduces the dataset size and thus computational time. The second, and arguably most important reason, is that as the goal of the scRNA-seq analysis is to identify cell types, and therefore keeping genes which are expressed at similar levels in all the cells (i.e. lowly variable genes) will only serve to make all the cells seem similar and thus, make them harder to separate. The most highly variable genes, however, will likely separate the cells in the dataset and reveal the underlying transcriptomic differences between the cells.

While identifying the top n highly variable genes helps to capture genes which are biologically meaningful, as the n is usually in the low thousands, there is still the issue of how to display the transcriptomic

relationship between cells in the dataset in a way which can be easily interpreted and how to further reduce the dimensionality of the data in order to save on computing cost and time. For this task, Principal Component Analysis (PCA) (Pearson 1901) is performed on the dataset, allowing the dimensionality of the dataset to be reduced, while maintaining some of the relevant signal in the high dimensional original data space. PCA requires that the data space has a mean of 0 and a standard deviation of 1, thus scaling of the data is performed to transform the data to fit these specifications (Fig.1F). Around the scaling step, many scRNA-seq data analysis tools allow users to regress the effect of specific variables out of one's dataset.

Intuitively, PCA involves fitting a line of best fit to the high dimensional data space and embeds the points (in the case of scRNA-seq data, these are cells) onto that line (Fig.1G). This line is the first Principal Component (PC), and constitutes a dimensionality reduction of the original data space down to one dimension. To create the second PC, a line orthogonal to the first PC is drawn through the data and, for the third PC, a line is drawn that is orthogonal to that of the second PC and so forth. Each subsequent PC will explain less and less variation of the original space, thus containing less biological information and more noise. There is thus a point at which including more PCs only adds unwanted noise into your analysis (Cattell 1966). To identify how many PCs to use, the proportion of variance captured by each PC is plotted, with the PCs that capture a high proportion of variation being used in future processes.

One of the main goals of scRNA-seq analysis is to be able to group cells in the dataset based on their transcriptome, in order to identify groups of similar types of cells. Groups of cells are identified using clustering methods, an umbrella term to describe methods which aims to group together individuals who have similar feature values (Fig.1H). Each clustering method has its own process for identifying clusters but, for scRNA-seq analysis, the two most popular clustering methods, developed by Leiden & Louvain (Traag, Waltman, and Eck 2019 & Blondel et al. 2008), have similar starting requirements for clustering. Using the cell embeddings in the PCA space, the k (by default Scanpy sets $k = 15$ and Seurat sets $k = 20$) nearest cells for each cell, in terms of euclidean distance, is stored in a graph. Because each cell has to be connected to k cells, cells that are isolated can end up connected to cells very far away from itself. As such, the graph needs to be pruned to remove these spurious connections, which can be achieved in different ways. In Seurat, Jaccard similarity is used to modify the edge weights. In the context of scRNA-seq data analysis, Jaccard similarity compares the number of shared nearest neighbours of two cells with the total number of nearest neighbours that the two cells have. Cells which share many nearest neighbours will have a high Jaccard similarity and are thus very similar to one another, while those that share few nearest neighbours will have the opposite. Thus, high Jaccard similarity positively weights the edges of similar (or close) cells, while having a negative effect on the weights of dissimilar (or far-away) cells. This refined nearest neighbours graph is the basis for Louvain and Leiden clustering.

Once clusters are created, the gene expression profiles of the cells within the clusters, and differences between the clusters, can be identified. However, before this step, further dimensionality reduction is usually carried out, but this represents a controversial area of transcriptomic analysis for a variety of reasons. Arguably the most widely used methods of further dimensionality reduction are Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018) and t-distributed Stochastic Neighbor Embedding (tSNE) (Maaten and Hinton 2008) (Fig.1I). The main difference between tSNE and UMAP is their preservation of local and/or global distances. Preservation of local distance can be thought of as grouping points together that are of the same type. Preservation of global distances, however, can be thought of as placing points near each other that are not identical, but are still similar in some way. As an example, let's take the fashion Modified National Institute of Standards and Technology database (fMNIST) dataset (Xiao, Rasul, and Vollgraf 2017), which contains black and white photos of various

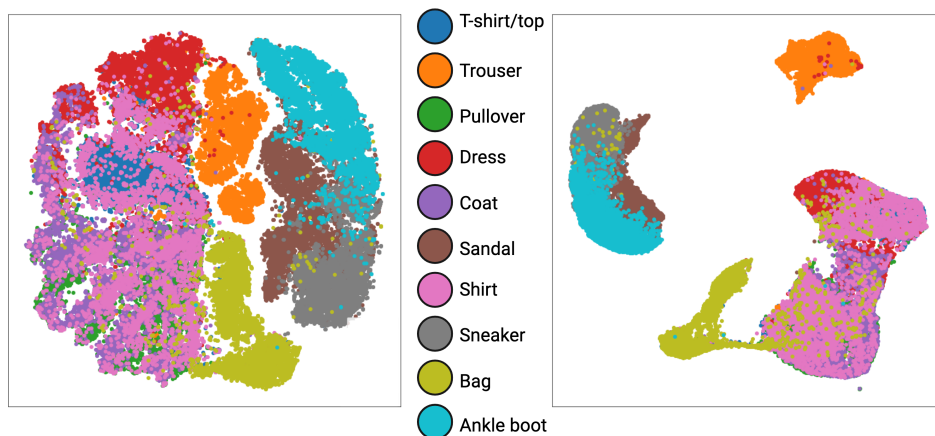


Figure 3: **fMINST dataset plotted using tSNE and UMAP**

fMINST dataset plotted using tSNE (left) and UMAP (right) embeddings. Points are coloured by what article of clothing they are.

articles of clothing. In figure 3, we can see the tSNE and UMAP embeddings of the fMINST dataset. On both embeddings, all images of the same type of clothing are close together; showing a preservation of local distances. When looking outside of the same type of clothing, on the UMAP, the footwear is spatially localised together away from everything else, while on the tSNE plot it is spatially adjacent to many other types of clothing. This shows a preservation of global distances by UMAP, but not by tSNE.

The claims that UMAP and tSNE can "preserve distances" are called into question by Chari & Pachter (Chari and Pachter 2023), who show that nearest cells in the ambient space (matrix of the normalised expression of the highest variable genes for all the cells) do not match the nearest cells in the UMAP or tSNE space, meaning local distances are not preserved. For global distances, the authors find that the list of closest cell types to a given cell type are often not well correlated between the UMAP/tSNE embeddings and the ambient space, suggesting that global distances are not well preserved either. This finding is not as impactful, as it assumes that any deviation from the ambient is equally wrong, when in reality some rankings are more wrong than others. To go back to the fMINST example, the ambient space closest types of clothing for sneakers being ankle boot, sandal then shirt and in the UMAP space being sandal, ankle boot then shirt is equally wrong (in the eyes of the metric) as the ambient space being ankle boot, sandals, shirt and the UMAP space being ankle boot, shirt, sandals (both have a Kendall's Tau = 0.333), despite the fact that thinking shirts and sandals are equally close to sneakers is incorrect (Fig.4).

Dimensionality reduction will corrupt the distances between cells due to the process of trying to project 1000s of dimensions worth of data into 2D. This will ultimately cause lack of preservation of original distances between the ambient and UMAP/tSNE space. The authors main issue with UMAP and tSNE is not that this distortion occurs, it is that they are being used without the knowledge that the distances are not necessarily preserved, and thus could bias conclusions which are based off of quantitatively interpreting the distances of cells (or cell types) on a UMAP or tSNE plot.

The final point in the pipeline, involves assigning a cell type/state identity to the clusters. The













Closest items of clothing to sneakers					
	Ambient 1	UMAP 1	Ambient 2	UMAP 2	
Closest item					
2nd closest item					
3rd closest item					
Kendall's Tau	0.33		0.33		

Figure 4: **Table showing a hypothetical ordering of similar items of clothing to sneakers**

Table detailing two hypothetical scenarios (blue and red) showing how close a picture of ankle boots, sandals and shirts are to sneakers in an ambient and UMAP space, with the Kendall's Tau correlation of the two scenarios displayed in the bottom row of the table.

most common means of annotating clusters is through marker gene analysis informed manual analysis or automated annotation with reference mapping tools (Fig.1J). Marker gene analysis involves comparing the gene expression across cells in one cluster with the expression in all the other cells in the dataset. Summary statistics are generated which include p-value, \log_2 fold change (\log_2 FC) and percentage of cells in which the gene is expressed. These statistics are used to filter the list of genes down to the markers of each of the clusters. Through assessing the established literature for these marker genes, cell types are then annotated. The issues with this approach are that it can be time consuming to search the literature for genes, and the top marker genes of a cell are not necessarily helpful in annotating the cell type. For example, the top marker genes of a cell that is undergoing proliferation will likely be cell cycle associated genes that, while informative, will not inform on what the cell type is. The advantage of reference mapping (Ianevski, Giri, and Aittokallio 2022 & Aran et al. 2019) is that it is automated and thus less time consuming than manually annotating the clusters. The downsides of this approach are the requirement of a reference (which may not exist) to use for mapping and the performance of the tools varying across different dataset contexts (Cheng et al. 2023).

5.3 Integrating scRNA-seq datasets

The pipeline described previously, and seen in Figure 1, applies to the scenario where one only requires a single dataset and a single condition to be analysed. It is commonplace to have many different conditions and controls to be compared; whether that be comparing wild type vs genetic perturbations, disease vs

healthy, or treated vs untreated. In any experiment, the aim should be to keep as many of the variables the same wherever possible. In a gel electrophoresis experiment, for example, one mitigates many biases by running all the samples on the same gel, at the same time, in the same place, at the same ambient temperature, with the same reagents on the same machine. In scRNA-seq, these variables rarely ever match up for a variety of reasons. For 10X based scRNA-seq, the more cells loaded means more doublets will be in the data (Genomics 2018), meaning large samples (or indeed numerous samples) will be split up into different runs. This fact means that several of the variables mentioned early are now different, all of which has an effect on the gene expression data that comes out at the end (Tran et al. 2020), making replicate samples seem different from one another and obscuring real biological variation. These technical differences are known as batch effects (Tung et al. 2017). To try and correct the data for these technical differences, integration of scRNA-seq data has become a standard part of the scRNA-seq analysis pipeline, when multiple datasets are compared.

As of November 2020, there were 49 scRNA-seq integration methods (Luecken, Buttner, et al. 2022), all of which have different approaches to the problem of integrating datasets. One of the most popular and best performing methods, Harmony (Korsunsky et al. 2019), creates clusters (global clusters) of cells on the PCA space which maximize diversity of batch. Within each cluster, groups batch into clusters (batch clusters) and calculate a correction factor, which brings the batch cluster centroids closer to the centroid of their assigned global cluster. From this, Harmony can then adjust the cell positions in the PCA space. Another method, batch balanced k nearest neighbours (BBKNN) (Polanski et al. 2020), constructs a k-nearest neighbours (kNN) graph of the PCA space where each cell is connected not to cells within its own batch, but to ones in the other batches. This kNN graph is then fed into a dimensionality reduction method such as UMAP to get the integrated embeddings of the cells. Given the variability in how integration methods correct for technical batch effects, comparison of integration method performance and consistency will be investigated throughout this thesis.

The strength of integration lies in its ability to allow fair comparisons between conditions, facilitating the identification of differentially expressed genes between the conditions for a given cell type, allowing researchers insight into what transcriptomic changes the conditions might cause to a type of cell (Briggs, Rojas, et al. 2021, Dooley et al. 2023 & Alivernini et al. 2020). There are, however, several issues which hinder the application of integration methods. The first issue is how beneficial the benchmarking metrics of integration quality are. One of the most commonly used metrics for comparing how good integration methods are is to see how much cell label overlap there is within clusters generated from the integrated space (Luecken, Buttner, et al. 2022). As benchmarking papers take a variety of different datasets from different papers, there is likely to be contradictions in annotation between them, thus standardisation needs to be carried out to ensure metrics can be performed as expected. This can, and has, lead to disagreement between the authors of the paper describing the data and the authors of the benchmarking paper as to what cell types are present in the data. For example, in the benchmarking paper by Luecken and colleagues (Luecken, Buttner, et al. 2022) they identify populations of Natural Killer (NK) T cells in datasets from Oetjen and colleagues (Oetjen et al. 2018) and from Sun and colleagues (Sun et al. 2019). The authors of the original papers, however, did not annotate any clusters as NK T cells. NK T cells are a type of invariant T cell which share characteristics of both T cells and NK cells, and are a relatively rare cell type (Maas-Bauer et al. 2024), making up only 0.19% of the immune cells in Tabula Sapiens dataset (Tabula Sapiens et al. 2022). In the human immune cell dataset used for benchmarking in the Luecken and colleagues paper, NK T cells make up 8% of the total cells and are more populous than the transcriptionally similar and more common CD8+ T cells, which make up 6% of the dataset. This disparity and uncertainty of cell type annotation will ultimately lead to the clustering derived metrics, assessing integration performance, to be biased and inaccurate.

The second issue is how to best tune an integration. In the Seurat integration tutorial (Seurat 2023), the integration requires one to choose the number of principal components one would like to use to integrate the data, but no guidance is given in the tutorial for choosing how many dimensions one should use. In Harmony, a convergence plot can be viewed to see how well the data has integrated towards a local minimum on the objective function. This plot, however, does not accurately reflect how optimally integrated the data is in some cases (Korsunsky 2020). Harmony also has other tuneable parameters with little solid guidance on how to tune them, such as its theta parameter, which controls how diverse the clusters are in terms of batch.

In a similar vein to lack of clarity on how to tune integration, there is a lack of clarity on how integration fits into a standard scRNA-seq pipeline. To illustrate this, reference will be made to the Scanpy tutorials for standard scRNA-seq analysis (Scanpy 2024) and the integration analysis (scanpy n.d.). At time of writing, the recommended standard scRNA-seq analysis pipeline tells users to scale their normalised count data by the median of total UMI counts over the cell, rather than the constant 10,000 which is often used. The integration tutorial starts off by loading in two separate datasets, with no normalisation steps being carried out before the integration takes place. As the datasets were loaded in separately, this gives users the, not unreasonable, assumption that datasets were normalised individually, before integration, and they would thus follow the Scanpy tutorial which details normalisation, with each dataset's normalised expression values being scaled by the median of its total counts. For datasets with the same median total counts, this will not matter, but often one dataset will have a higher median total counts and thus their expression values will be inflated relative to the other dataset, affecting possibly effecting differentially expressed (DE) gene analysis. This shows how the lack of clarity not only affects integration, but can also affect all the downstream processes of it.

Despite these downsides, the process of integrating data is an essential part of scRNA-seq analysis and more research should be devoted to this subject to ensure that it is continued to be investigated and improved upon.

5.4 Trajectory inference

Much of the previous sections has focused on the preprocessing and integration steps that are required to make scRNA-seq data analysable in a meaningful way. Now, the focus will turn to the methods that can be applied downstream of these preprocessing steps to generate results and discover novel biology.

Many systems of biology are continuous processes that occur over a span of time. Examples of such include parasite life cycles (Briggs, Rojas, et al. 2021), haematopoiesis (H. Chen et al. 2019), embryogenesis (Trapnell et al. 2014), immune cell activation, to name but a few. As the transcriptomes of the cells undergoing these processes will change across time to facilitate progression through the underlying biological process, the transcriptomic changes that occur across the biological process will be encoded in the scRNA-seq data. By accessing this data, more information on the transcriptomic changes that occur, and even drive, these processes could be better understood. The field of scRNA-seq analysis which deals with extracting gene expression changes overtime is Trajectory Inference (TI). Like integration, the specifics of how each method derives a trajectory depends on the method, but at its simplest level, most TI methods can be thought of as a join-the-dot puzzle solved with mathematics, rather than the human eye.

Most TI methods first require cells to be projected in some low dimensional space. Tools range from being apathetic about what reduced dimension space is used (e.g. Slingshot (Street et al. 2018)) to recommending one particular method like Monocle (Trapnell et al. 2014, X. Qiu et al. 2017 & Cao et al.

2019), which runs based on UMAP embeddings. There are even some dimensionality reduction methods that have been built with the express purpose of capturing scRNA-seq trajectories like PHATE (Moon et al. 2019) (Fig.5A).

Given that the embedding of cells in a low dimension space can have outliers or differences in distribution patterns, it would not be computationally reasonable, or indeed sensible, to construct a trajectory connecting each cell in the dataset to one another. Instead, most TI methods create clusters, or take user-defined clusters for the dataset and find the centroids of these clusters (Fig.5B). A tree is then constructed between the centroids in such a way that minimises the distances between the centroids (Fig.5C). As the tree is undirected, TI cannot give any guidance on where the start and end points of the process are. Therefore, TI methods asks the users to define the start point of the trajectory, and identify the possible paths from the start centroid to the end centroids, which the methods define as centroids which only have one edge. Cells are then projected onto the closest point of the tree. This can be thought of as a projection onto a one dimensional line (Fig.5D).

From the projection of cells onto the line, the approximation of how far each cell is through the biological process (referred to as pseudotime) can be calculated. Pseudotime values are assigned in a variety of ways, but usually a starting cell is chosen and then the closest cell to that has its pseudotime set equal to the euclidean distance from it to the starting cell. The closest point to that cell then has its pseudotime assigned as the pseudotime of the cell, plus the euclidean distance of the closest cell to it. This process repeats until every cell is assigned a pseudotime value (Fig.5E). Through assigning cell pseudotime, it's possible to look at gene expression changes over time.

TI is not without its downsides. As dimensionality reduction methods play a key role in determining the trajectory output, the issues surrounding the distortion of cell relationships when reducing dimensionality will plague the output of TI. Another issue is the reliance on the user to choose the start and end points of the trajectory. For areas of biology which have been studied, this of course could be derived from the literature. In the scenarios where the start and end points are unknown, TI is unable to identify the order of the gene expression changes that occur across the process. This could lead to inaccurate characterisation of the biological process; especially as TI packages often randomly choose start and end points when the user does not define them. This problem can be ameliorated with the use of RNA velocity, which utilises the ratios of unspliced to spliced mRNA to calculate the initial (high rates of transcription, thus high unspliced mRNA) and terminal (high rates of splicing and degradation, thus low unspliced mRNA) states of a biological process and thus the order of cells through a process can be recapitulated (La Manno et al. 2018). While RNA velocity can be combined with TI to help produce more biologically accurate and meaningful results (Lange et al. 2022), its binary decision of unspliced vs spliced mRNA and reliance on dimensionality reduction methods for visualisation (among other things) means that its overall accuracy has been called into question (Gorin et al. 2022).

Both TI and RNA velocity are indirect estimates of cell fates and movements through their development. To perform more accurate analysis of cell developmental processes and fates, direct ways of inferring these are required. Such approaches include metabolic labelling and lineage tracing. The former process labels nascent RNA with 4sU, which can be converted into a cytosine analog when induced with iodoacetamide. By looking at the detection of U to C mutations in the RNA compared to the reference, old vs new RNA transcripts can be distinguished, and thus a cells velocity estimated. RNA velocity uses spliced to unspliced RNA as a proxy of cell velocity, but the proportion of unspliced reads can be very low (Gorin et al. 2022) and the percentage of unspliced reads has high variability between genes (Maizels, Snell, and Briscoe 2024), making estimation of nascent transcription prone to noise. The distribution of metabolic labelled RNA is higher for genes compared with unspliced reads, with a more

uniform distribution across genes, meaning metabolic labelling provides less noisy estimates of RNA velocity dynamics (??). Lineage tracing on the other hand allows cell lineages to be traced through an expressible, heritable barcode which can be induced to mutate its sequence. The changes in the expressible barcode over cell generations allows daughter cells to be identified, and thus a cell lineages traced (Wagner and A. M. Klein 2020).

Despite its downsides, TI has been applied in a variety of contexts to reveal novel biology. One particular area of application is to analyse the life cycle of parasites (Briggs, Rojas, et al. 2021, Real et al. 2021, Howick, Russell, et al. 2019 & Howick, L. Peacock, et al. 2022). Parasite life cycles lends itself to TI analysis as, unlike mammalian cell samples, which could contain a variety of cells undergoing independent biological processes, the transition through the life cycle stages is the central process that drives transmission and infection.

5.5 ScRNA-seq analysis of parasites

The earliest paper to apply scRNA-seq to parasites was by Poran and colleagues (Poran et al. 2017), who used a droplet-based UMI scRNA-seq method to profile the transcriptomic changes that accompany sexual commitment in *Plasmodium falciparum*. In this paper they showed that despite the parasite having less RNA content than mammalian cells, the average number of transcripts recovered per cell was similar to that of mammalian cells (Poran et al. 2017). They also identify several novel genes which may facilitate sexual commitment in the parasite, showing that scRNA-seq is a powerful tool for investigating the transcriptomic changes that occur across the parasite life cycle. Since the Poran and colleagues paper, many more researchers have used scRNA-seq to profile parasites. In 2019 the first single cell transcriptomic atlas of a parasites life cycle stages was generated by Howick and colleagues (Howick, Russell, et al. 2019) for *Plasmodium falciparum*, identifying marker genes for all its life cycle stages. Application of scRNA-seq has spread out to other parasites like *Schistosoma mansoni* (Wendt et al. 2020 & Diaz Soria, Lee, et al. 2020) and *Toxoplasma* (Xue et al. 2020). In 2018, the first scRNA-seq analysis of a member of the kinetoplastid group, *Trypanosoma brucei* (*T. brucei*), was carried out by Müller and colleagues (Müller et al. 2018), showing the expression of VSG across WT and histone variant knockout BSF and metacyclic cells. Since then, the application of scRNA-seq method to *T. brucei* has continued, shedding light on the various transitions and stages occurring across its life cycle (Vigneron et al. 2020, Briggs, Rojas, et al. 2021, Hutchinson et al. 2021, Howick, L. Peacock, et al. 2022 & Briggs, Marques, et al. 2023); opening opportunities to apply this method to study other members of the kinetoplastids.

Despite the existence of multiple papers which apply scRNA-seq to parasites, the presence of parasite datasets in the papers describing the scRNA-seq analysis methods themselves (or subsequent benchmarks) is rare. For example, the original papers describing the 15 integration methods that Luecken and colleagues (Luecken, Buttner, et al. 2022) test in figure 3 of their benchmark paper, do no contain any applications to parasite scRNA-seq datasets. Furthermore in the benchmark paper, itself only human and mouse datasets are tested. There is thus a lack of information on how these methods perform on parasite datasets in general.

5.6 Introduction to Kinetoplastids

The kinetoplastids comprise a group of protists characterised by the presence of a kinetoplast; a structure within the parasite mitochondrion which contains encodes for a variety of genes (Amodeo, Bregy, and Ochsenreiter 2023). Some species within the kinetoplastid group contribute to a variety of human and animal diseases: *T. brucei* which causes human and animal African Trypanosomiasis, *Trypanosoma cruzi* (*T. cruzi*) which causes Chagas disease, *Leishmania* which causes Leishmaniasis and a variety of

Trypanosomes which cause the animal specific disease Nagana (K. Stuart et al. 2008 & Fetene et al. 2021).

The gene expression of kinetoplastids is polycistronic (Clayton 2019), meaning multiple genes are co-transcribed at the same time as one continuous polycistron unit and later undergo trans-splicing into individual transcripts. The consequence of this is that gene expression regulation is not facilitated through controlled activation of RNA II polymerase activity, rather through controlling post-transcriptional processes including mRNA stability and export (Fernandez-Moya and Estevez 2010). This makes mRNA binding factors an area of great focus in kinetoplastid research, as they are important facilitators of gene regulation and thus play a key role in regulating essential biological processes within the parasite. Many mRNA binding factors that affect mRNA stability in *T. brucei* have been identified (Archer et al. 2009, Hartmann et al. 2007 & Ling, Trotter, and Hendriks 2011), with each being associated with different biological functions such as cell cycle (Archer et al. 2009) or stabilising life cycle stage-specific transcripts (Ling, Trotter, and Hendriks 2011).

5.7 The life cycle of *T. brucei*

Arguably the most well-studied of the kinetoplastid family is *Trypanosoma brucei* (*T. brucei*), given (among other things) the stability of culture *in vitro* (Cunningham 1977, Baltz et al. 1985 & H. Hirumi and K. Hirumi 1989), capacity to carry out CRISPR/Cas9 gene editing (Rico, Jeacock, et al. 2018), ability to induce overexpression of genes and perform RNAi, the latter not being possible in many *Leishmania* subgenus (Lye et al. 2022) or in *T. cruzi* (Kolev, Tschudi, and Ullu 2011).

The life cycle of *T. brucei* takes place over two main phases, a mammalian and insect stage (Matthews 2005). The first major form in the mammalian stage of the parasite life cycle is the metacyclics. These are a cell cycle-arrested form of the parasite which is preadapted for the mammalian host environment upon transmission during a bloodmeal due to the upregulation of glycolysis associated genes and proteins and expression of surface metacyclic variable surface glycoprotein (mVSG) (Christiano et al. 2017). Upon transmission into the mammal, the metacyclics rapidly differentiate into slender form cells (Matthews 2005). This form readily replicates within the host; protected from the host's innate and adaptive immune system by a Variable Surface Glycoprotein (VSG) coat (Barry and McCulloch 2001). The slender forms have the capacity to infect teneral tsetse flies (the parasites vector) upon bloodmeal (Schuster et al. 2021), however the efficiency of infection is very low, with the brunt of successful infections being achieved by the next mammal stage in *T. brucei* development: the cell cycle arrested stumpy form (Matthews, Ellis, and Paterou 2004 & Ngoune et al. 2024), which are pre-adapted to infect the teneral tsetse fly (Rico, Rojas, et al. 2013). Stumpy forms develop from slender forms through a quorum sensing mechanism, triggered by high cell density (Reuner et al. 1997) and driven by the accumulation of oligopeptides, which are transported by a G protein-coupled receptor encoded by *TbGPR89* into the cell to induce the differentiation of slender to stumpy forms (Rojas et al. 2019). These oligopeptides are generated from host peptides by peptidases secreted by the parasite into the environment (Tetty, Rojas, and Matthews 2022). Upon taking a bloodmeal from an infected host, stumpy forms are transferred into the teneral tsetse fly where they differentiate into procyclic forms (PCF) in the posterior midgut of the fly (Rotureau and Van Den Abbeele 2013).

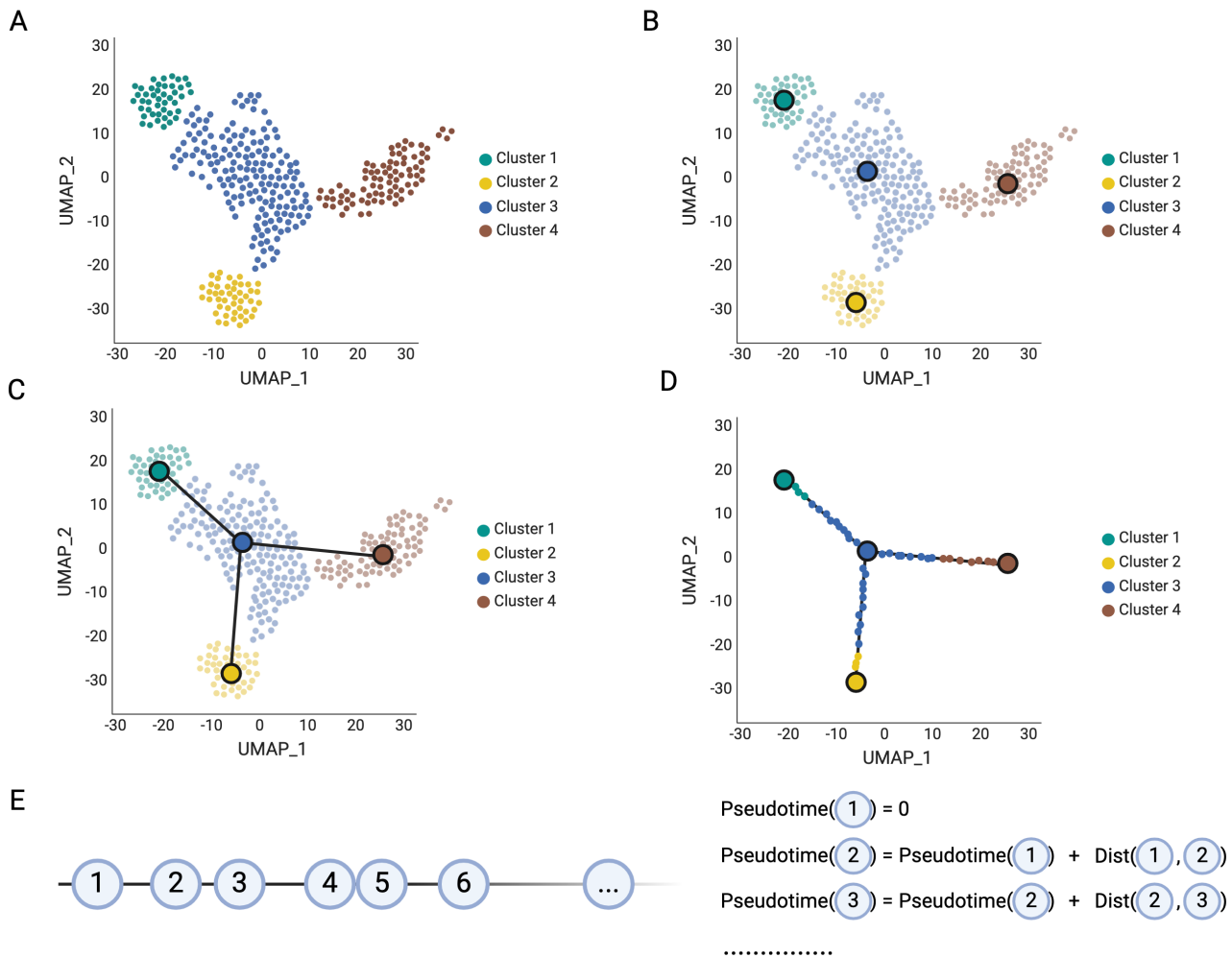


Figure 5: **Overview of the Trajectory Inference process**

A typical Trajectory Inference (TI) process starts with projecting the cells into a reduced dimension space (e.g. UMAP) and assigning cells into clusters (A). Centroids are then calculated for the clusters (B) and a tree is constructed using the cluster centroids as nodes in the tree (C). Cells are then projected onto the closest point of the tree line (D) and pseudotime is calculated by summing the pseudotime of the previous cell and the distance (E).

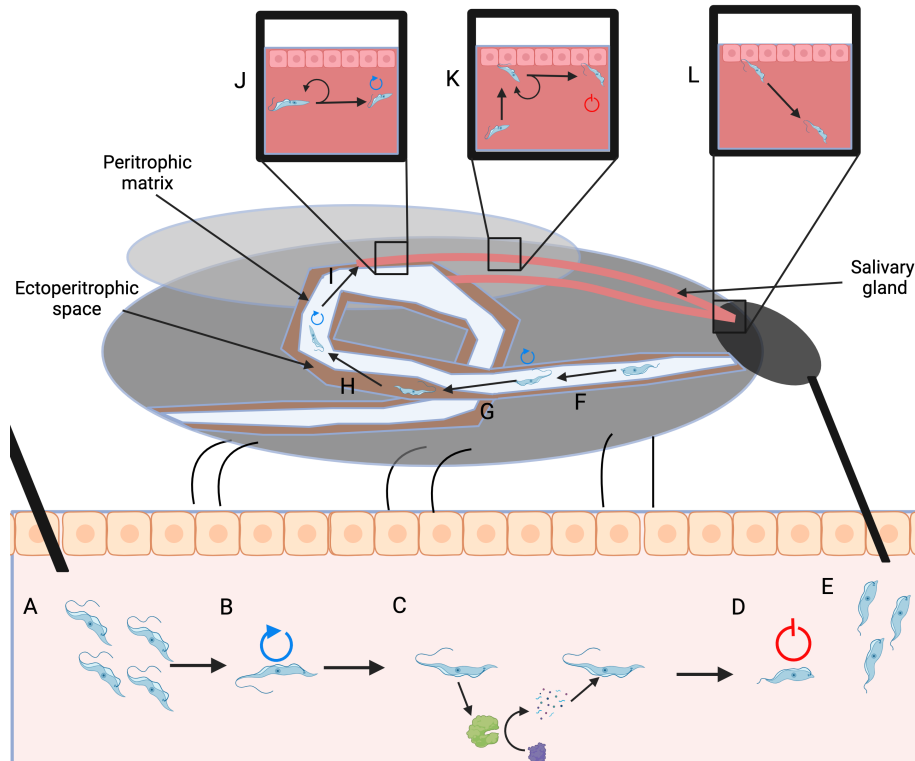


Figure 6: Artistic impression of the *T. brucei* life cycle

Metacyclic form parasites are transmitted into the mammalian host when a bloodmeal is taken by an infected tsetse fly (A). The metacyclic forms then transition into replicative slender forms (B). The slender forms release peptidases which break down host proteins into oligopeptides, which are then sensed by the parasites (C). The accumulating oligopeptides eventually signal the slender parasite to transition into its non-replicative stumpy form (D). The stumpy forms are then taken up into the tsetse fly during a blood meal (E). The stumpy forms transition into replicative PCFs (F) before crossing the peritrophic matrix into the ectoperitrophic space (G). PCFs eventually cross back through the peritrophic matrix into the gut lumen and transition into replicative long epimastigotes (H). The long epimastigotes subsequently invade the salivary glands of the parasite (I) and divide asymmetrically into long and short epimastigotes (J). The short epimastigotes attach to the salivary gland epithelium and colonise it, before differentiating into mammal infective, non-replicative metacyclics (K). Metacyclics detach from the epithelium and invade the host upon bloodmeal of the parasite, restarting the life cycle (L). This figure is partially adapted from Aksoy 2019 (S. Aksoy 2019). Replicative parasite forms are denoted by the blue circle arrow and non-replicative forms are denoted by the red barred circle.

When the BSF parasite is exposed to a coldshock *in vitro*, it causes the expression of PCF-associated surface protein EP procyclin and makes the parasites more sensitive to citrate/cis aconitate (CCA) (Engstler and Boshart 2004), which is used to differentiate bloodstream forms into PCF *in vitro* (Czichos, Nonnengaesser, and Ovrath 1986 & Ziegelbauer et al. 1990). Coldshock also induces the trafficking of Protein Associated with Differentiation (PAD) 2 to the surface of stumpy forms which belongs to a family of proteins that help mediate stumpy to PCF transition, as the inhibition of their gene expression through RNA interference causes a reduction in CCA responsiveness and PCF development (Dean et al. 2009).

In the posterior midgut, the PCFs divide readily and shed their VSG coat (Rotureau and Van Den Abbeele 2013). The shed VSG can be taken up by tsetse fly cardia cells, which maintain the midgut peritrophic matrix (PM), and causes changes in these cells, like the downregulation of a microRNA mir-275, disrupting the regular barrier capabilities of the PM (E. Aksoy et al. 2016), possibly aiding the parasite to move across the PM barrier. The PCFs cross into the anterior midgut and differentiate into the elongated, cell cycle-arrested mesocyclic stages, which subsequently migrate into the proventriculus, where they differentiate into the replicative epimastigotes (Van Den Abbeele et al. 1999 & Rotureau and Van Den Abbeele 2013). The epimastigotes migrate into and along the salivary glands (SG), dividing asymmetrically into a long and short form (Van Den Abbeele et al. 1999) with it hypothesised that the short forms play the main role in attaching to, and subsequently colonising, the SG epithelium (Oberle et al. 2010). The attached epimastigotes then divide asymmetrically again, giving rise to more short-form epimastigotes, as well as metacyclic cells; restarting the life cycle upon transmission into mammals (Rotureau, Subota, et al. 2012).

5.8 The life cycle of *T. cruzi*

The life cycle of *T. cruzi*, like *T. brucei*, is split across a mammalian stage and an insect stage, in this case members of the triatominae family. When triatomines penetrate the skin of a mammal during a bloodmeal, metacyclic trypomastigotes (contained within the faeces and urine of the triatomine) can pass into the host and subsequently invade the host cells (Martin-Escolano et al. 2022). Between crossing into the host tissue and entering the host cells, the parasite is able to be targeted by many components of the host immune system. To allow survival in the host, the parasite has evolved a variety of mechanisms to evade the immune system. One of the key defense mechanisms *T. cruzi* has is to subvert a mechanism the immune system uses to distinguish self from non-self, namely the presence of sialic acid (SIA). SIA is found on the surface of all mammalian cells and helps protect itself from immune attack through a variety of mechanisms, an example being inhibiting the binding of complement to the cell surface (Varki 2011). *T. cruzi* is susceptible to such an attack, but it is able to effectively steal host cell surface-bound SIA through the action of trans-sialidases, a highly polymorphic family of genes which encode surface proteins that cleave SIA off host cells and transfer them to the surface of the parasite, making them resistant to attacks from the complement system (Fonseca et al. 2019 & Nardy, Freire-de-Lima, and Morrot 2015).

While the parasite is protected from the host immune system, to progress to the next life cycle stage, it needs to invade a host cell. The invasion mechanism of the parasite has not been fully elucidated but there are known interactions that occur between the parasite and host cell (Rodriguez-Bejarano, Avendano, and Patarroyo 2021). The parasite has a variety of host proteins which facilitate interactions between the parasite and host, and have been shown to be involved in parasite invasion of host cells. The glycoprotein 82 (gp82) is one such protein, with research showing that the binding of this protein by antibodies significantly reduces the infection capabilities of metacyclic trypomastigotes (Ramirez et al. 1993). Gp82 binding to host cells causes the release of calcium ions into the cytosol of the host and para-

site (Ruiz et al. 1998), through the activation of phospholipase C (Yoshida et al. 2000). This increase in calcium ion concentration leads to exocytosis of lysosomes (Martins et al. 2011), which promotes uptake of the parasite into a parasitophorous vacuole (PV) (Batista et al. 2020). Unlike in *Leishmania*, *T. cruzi* does not persist inside the PV, rather it escapes into the cytosol to continue its life cycle (Batista et al. 2020). This egress of the parasite from the PV into the cytosol is dependent on a low pH environment, with the parasite egress being inhibited when the pH of the PV is raised (Ley et al. 1990). This low pH environment not only allows the transition of metacyclic trypomastigotes into the next life cycle stage, amastigotes (Tomlinson et al. 1995), but it also activates Tc-TOX, a protein secreted by the parasite which functions at a low pH. Tc-TOX has been shown to form pores in the membranes of the PV; allowing parasite egress from the PV (Andrews and Whitlow 1989 & Stecconi-Silva, Andreoli, and Mortara 2003). Within the cytosol, amastigotes begin to replicate and undergo a change in transcriptome that allows them to adapt to their intracellular environment. The expression of genes encoding surface proteins and flagellar components is downregulated, while a variety of metabolism processes like oxidative phosphorylation and fatty acid synthesis increase (Y. Li et al. 2016) due to parasite transport of host molecules like glucose, glutamine and triacylglycerols (Shah-Simpson et al. 2017 & Gazos-Lopes et al. 2017), which fuel their growth.

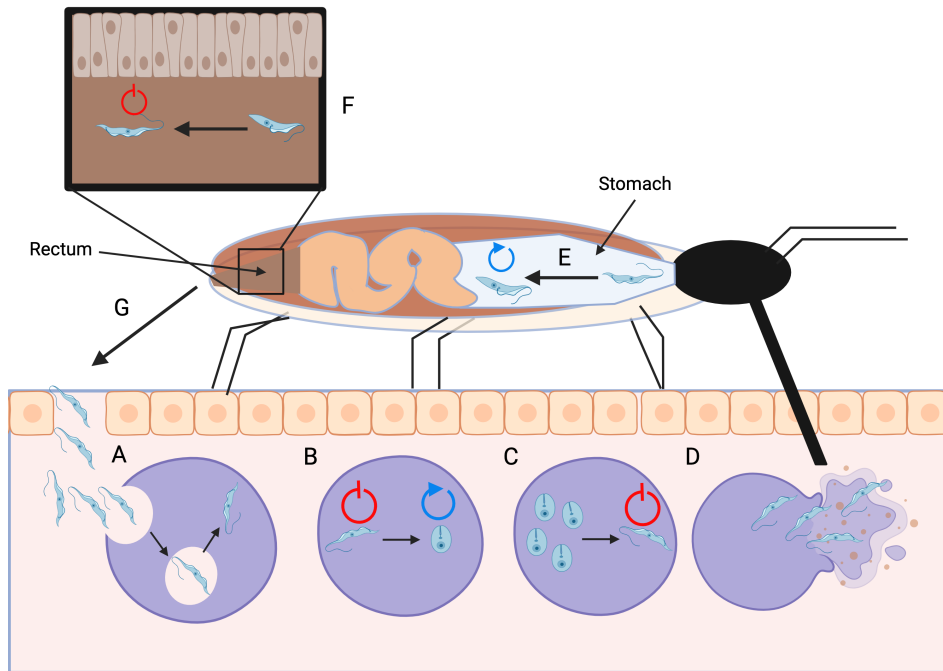


Figure 7: Artistic impression of the *T. cruzi* life cycle

Metacyclic trypomastigote forms enter the mammalian host through a wound created when the triatomine takes a bloodmeal, where they infect host cells through uptake into a parasitophorous vacuole, which they subsequently escape from (A). The metacyclic trypomastigotes then transition into replicative amastigotes (B) which divide readily before transitioning into non-replicative trypomastigote forms (C). The trypomastigote forms then egress from the cell, where they are taken up by the triatomine during a bloodmeal (D). In the triatomine's stomach, the trypomastigotes differentiate into replicative epimastigote forms (E) which then transition through the triatomine to the rectum where they differentiate into non-replicative metacyclic trypomastigotes (F) which are then excreted through the rectum and can infect the mammalian host, restarting the life cycle (G). Replicative parasite forms are denoted by the blue circle arrow and non-replicative forms are denoted by the red barred circle.

Figure is partially adapted from Garcia and colleagues (E. S. Garcia et al. 2010).

The parasites multiply inside the host cells and, at some point, begin their differentiation into trypomastigote forms. These forms are non-replicative (Taylor et al. 2020) and need to be taken up by the triatomine to progress to the next stage in their life cycle, thus the parasite needs to egress from the host cell. For egress, the parasite causes changes in the cytoskeleton of the host cell, such as a decrease in filamentous actin deposition (Ferreira et al. 2021). These changes in tandem with, or perhaps caused by, protease activity (Low, Paulin, and Keith 1992) allow the cell to egress from the host cell through the rupturing of the host cell membrane.

During a bloodmeal, released trypomastigotes can be taken into the triatomine where they enter the midgut and differentiate into replicative epimastigote forms. The epimastigotes then migrate to the rectum of the triatomine, where they attach to rectal epithelium. Epimastigotes then differentiate into non-replicative metacyclic trypomastigotes to start the life cycle over (E. Garcia and Azambuja 1991). *In vitro*, metacyclogenesis can be induced by nutrient starvation, cyclic AMP analogs or adenylate cyclase activators (Gonzales-Perdomo, Romero, and Goldenberg 1988), but what induces metacyclogenesis *in vivo* is unknown (Goncalves et al. 2018).

5.9 Problems with scRNA-seq analysis of Trypanosomatids

A common thread when discussing Trypanosomatids is the breadth of knowledge on the mammalian forms of the parasite and the dearth of knowledge of the insect forms. For *T. brucei*, while bloodstream and PCF parasites can be readily cultured *in vitro* (H. Hirumi, Doyle, and K. Hirumi 1977 & Cross and Manning 1973), and the transitions from slender to stumpy and stumpy to PCF can be recapitulated *in vitro* (Rojas et al. 2019 & Czichos, Nonnengaesser, and Ovrath 1986), there is no currently published way to generate the PCF to epimastigote or epimastigote metacyclic transitions *in vitro* using wild type (WT) cells. Thus, the insect stages need to be derived from the tsetse fly itself, where successful infections can be difficult to accomplish (Ngoune et al. 2024). In contrast to *T. brucei*, all the life cycle transitions of *T. cruzi* are readily captured *in vitro* (Kessler et al. 2017, Contreras et al. 1985, Piras 1982 & Hashimoto et al. 2015), however the high category risk nature of the parasite makes working with *T. cruzi* difficult and thus no scRNA-seq data has been generated on the parasite.

The problems with scRNA-seq analysis of Trypanosomatids extend beyond issues with culturing and collecting the parasites, the data generated from these analyses is also difficult to utilise completely. A common step counted map reads is to remove reads that map to multiple places on the genome, as their exact origin in the genome cannot be accurately determined. In *T. brucei* data generated by Vigneron and colleagues (Vigneron et al. 2020) using 10X 3' V2 chemistry, 81% of reads mapped to the *T. brucei* TREU927 reference genome but the percentage of reads that mapped confidently to the transcriptome falls to 40.50%. In contrast, the percentage of reads mapping to genome for immune cells from the blood of a human (generated with the same chemistry kit as the *T. brucei* data) was 95.2%, with the percentage confidently mapping to the genome being 92.7% (Genomics 2017). The relatively increase in reads that map to multiple locations on the genome in *T. brucei* compared with human immune cells is likely due to the presence of multigene families which are often highly repetitive in sequence (Pita et al. 2019). The practical implications of this is the true gene counts for these repetitive genes could be underestimated and not a insignificant percentage of the reads captured are unusable for scRNA-seq analysis and thus is wasted sequencing. Due to the nature of gene expression in Trypanosomatids, some scRNA-seq analysis methods are not currently applicable. RNA velocity relies on the ratios of unspliced to spliced reads to calculate cell fates, but due to the polycistronic regulation of gene expression, such information is not available for Trypanosomatids, where only two genes have introns (Berriman et al. 2005 & Siegel et al. 2010).

6 Aims and Objectives

This thesis seeks to address several aims and objectives across three results chapters.

Chapter 1 - To create a tool to identify changes that occur across time between conditions, a tool (TrAGEDY) was made that aligns two scRNA-seq trajectories, creating a common time axis on which the cells can be compared and gene expressions difference identified between the processes.

Chapter 2 - To create the first scRNA-seq atlas of *in vitro* *T. cruzi* development and identify transcriptomic changes that define the *T. cruzi* life cycle stages and their transitions.

Chapter 3 - To investigate the gene expression changes which drive the insect stage *T. brucei* life cycle by analysing scRNA-seq data of the *RBP6* overexpression model of *T. brucei* insect life cycle stage development.

7 Chapter 1: TrAGEDy - Trajectory Alignment of Gene Expression Dynamics

The following chapter is based off of a paper written by myself which is available on BioRxiv: Laidlaw et al. 2024.

To compare biological processes between conditions, alignment of single cell transcriptomic trajectories can be performed. Current tools place constraints on what data can be aligned. In this chapter I present Trajectory Alignment of Gene Expression Dynamics (TrAGEDy) which overcomes these constraints, allowing the alignment of asymmetric biological processes. Across simulated and real datasets, TrAGEDy returns correct alignment based on underlying simulated process, where current tools fail. Across different biological contexts, TrAGEDy can capture biologically relevant genes which other differential expression methods fail to detect. TrAGEDy provides a new tool for in-depth analysis of many emerging single cell transcriptomic datasets.

8 Chapter 1: Introduction

First described in 2014 (Trapnell et al. 2014), Trajectory Inference (TI) methods order cells based on the gradual change in transcript levels in underlying biological processes captured by single cell RNA-sequencing (scRNA-seq). Cells are assigned a “pseudotime” value based on their position in the inferred trajectory, allowing differential gene expression tests over pseudotime. Various methods have been developed to identify differentially expressed (DE) genes in a biological process of interest, in some cases allowing targeted comparisons. Notably, TradeSeq (Van den Berge et al. 2020) uses negative binomial general additive models (GAM) and Wald tests to assess whether a gene is DE over pseudotime within a trajectory or between lineages of the same trajectory. Alternatively, Monocle’s BEAM (X. Qiu et al. 2017) identifies genes associated with a particular branch of a trajectory, whereas pseudotimeDE (Song and J. J. Li 2021) uses permutation to account for uncertainty in pseudotime and a zero-inflated negative binomial GAM to account for expression value dropout. By applying these techniques, dynamic gene expression patterns associated with a variety of interesting biological events have been investigated, such as regulators of myogenesis (Trapnell et al. 2014), effector gradients in CD4+ T cells (Cano-Gamez et al. 2020), and life cycle transitions of the pathogen *T. brucei* (Briggs, Rojas, et al. 2021).

Some of these methods can be extended to identify DE genes between two conditions where the common biological process under analysis varies, such as before and after mutation or in an overlapping response to different stimuli (Van den Berge et al. 2020). One approach is to integrate the separate datasets together and complete a cluster-based comparison between conditions, treating cells as being at a discrete stage in development captured in each cluster. A limitation of this approach, however, is that development may be continuous, with some cells falling between stage boundaries. Another approach is to integrate the datasets together, perform TI to capture the shared, possibly branching, trajectory and differential expression tests across this pseudotime axis for each condition, using tools like TradeSeq’s conditionTest. However, the integration may force similarity between the trajectories, thus obscuring DE genes.

Alternatively, trajectory alignment can align cells from independently generated trajectories to find a common pseudotime axis that retains the original ordering of cells without dataset integration. Analysing aligned trajectories can reveal DE genes across the captured process, as well as differences between conditions. The discrete point at which the genes are DE between conditions can also be found. The first method described to perform such trajectory alignment was cellAlign (Alpert et al. 2018), which uses dynamic time warping (DTW) to align two trajectories together. Interpolated points are created across the trajectory, and gene expression values are assigned to the interpolated points - based on the cells that are in pseudotime vicinity to the points. Similarity between the conditions is assessed by computing the distance or correlation between the scaled gene expression values for the interpolated points on opposing conditions and, from this, the minimum cost path through the interpolated points is calculated using DTW.

The use of DTW imposes some constraints on the type of alignment that can be analysed. Each trajectory being compared must have the same start and end points, and each part of a trajectory must be matched to at least one point in the other trajectory. The practical implication of such limitations is that cell types that may not be represented on the opposing trajectory are nevertheless matched to a point on the other trajectory, making interpretation of the subsequent alignment difficult.

Here, we build on the work of cellAlign with Trajectory Alignment of Gene Expression Dynamics (TrAGEDy), where we make post-hoc changes to the alignment, allowing us to overcome the limitations of DTW and better reflect differences that may occur across the alignment. We also implement an approach to identify DE genes across the alignment, in order to better identify differences between the two conditions. We test TrAGEDy with a variety of simulated scRNA-seq datasets with different underlying alignments, as well datasets of *T. brucei* life cycle and T cell development under different genetic conditions, revealing gene expression changes undetected by current methods without the need for prior data integration.

9 Chapter 1: Results

A typical workflow for TrAGEDy is as follows (Fig.8). Prior to trajectory alignment with TrAGEDy, a predefined list of marker genes of the various datasets being aligned is created; top variable genes or cluster markers can be selected, for example. The two datasets are then projected individually into a reduced dimension space using the preselected genes and TI performed on the two datasets independently (Fig.8A). There is no limitation on what TI method can be used, provided that pseudotime values are generated for each cell.

TrAGEDy then uses the cellAlign method of creating interpolated points to sample gene expression

at different points in the trajectory, allowing inherently noisy scRNA-seq data to be smoothed across the trajectory. The closer a cell is to the interpolated point, the more it contributes to its gene expression. By changing a parameter that controls a Gaussian pseudotime window, the user can alter the contribution level that more distal cells have to interpolated point gene expression (Fig.8B).

Dissimilarity between the gene expression of the interpolated points is next assessed. While interpolation can smooth out noise, possible batch effects between datasets mean calculating dissimilarity with methods such as Euclidean distance is problematic without scaling the data first. TrAGEDy uses Spearman correlation, as it is less sensitive to outliers than Pearson and, unlike Euclidean distance does not require scaling of gene expression values.

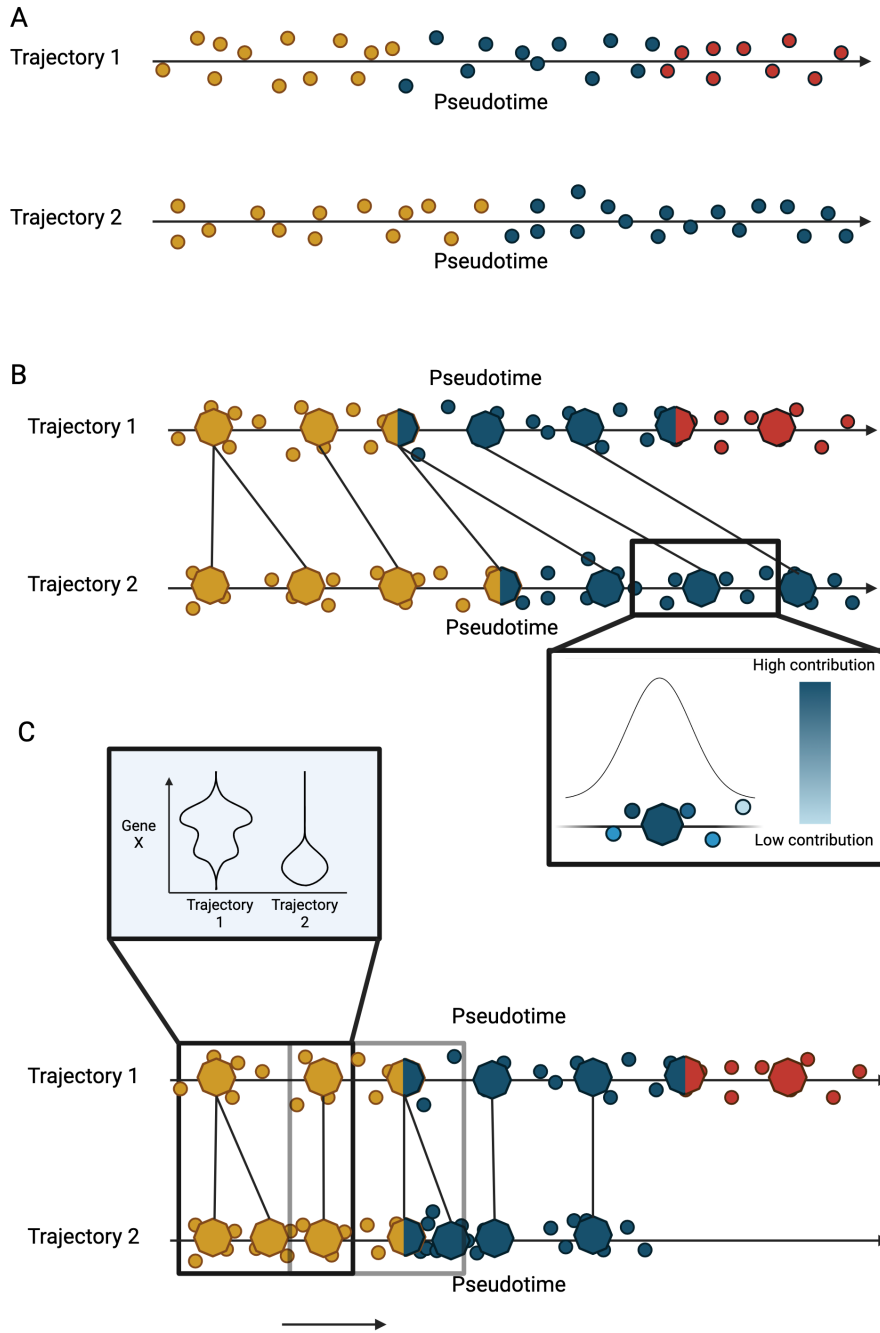


Figure 8: Graphical overview of the TrAGEDy process.

Trajectory Inference is carried out on two datasets that share a common process but have a difference in condition (A). TrAGEDy samples gene expression across interpolated points of the trajectory, with cells with closer pseudotime values to the interpolated points contributing more to their gene expression (B). TrAGEDy then aligns the pseudotime of the interpolated points then the cells, finally performing a sliding window comparison between cells at similar points in aligned pseudotime, thereby extracting DE genes (C).

TrAGEDy next finds the optimal path through the dissimilarity matrix of the interpolated points, which constitutes the shared process between the two trajectories. DTW, with alterations, is used to find the optimal path. To overcome the constraint in DTW that the two processes must start and end at the same point, we first cut the dissimilarity matrix so the path starts and ends at points that are most transcriptionally similar to one another (appendix Fig.2). Another constraint of DTW is that all points must be matched to at least one other point; thus (post-DTW), we pruned any matches that have high transcriptional dissimilarity, enabling processes that may have diverged in the middle of their respective trajectories to be compared (appendix Fig.3). Finally, we used the modified alignment to adjust the pseudotime values of the interpolated points, generating new values as the basis to then adjust the pseudotime value of each cell (Fig.8C).

To extract DE genes, we performed a sliding window comparison, soft clustering cells at similar points in aligned pseudotime together and then testing for significance with a Mann-Whitney U test, extracting genes that are DE at different points in the shared process (Fig.8C). Soft clustering with a sliding window allows comparisons over a continuous process. Users can define how much overlap there is between the windows of comparison.

9.1 Aligning simulated scRNA-seq datasets

We tested TrAGEDy’s ability to correctly identify the alignment of trajectories on datasets where the alignment points were known a priori. Using Dyngen (Cannoodt et al. 2021) we simulated scRNA-seq datasets with inherent trajectories encoded in the expression values, resulting in three different test cases: two positive controls and one negative control (Fig.9A).

The first positive test case used two datasets, the first simulating a full developmental progression involving sequential expression of many genes (wild type; WT), and the second resulting from knockout (KO) of one of these genes (Fig.9A). Thus, the KO trajectory is truncated at the point at which the process requires the transcription of the knocked out gene. TrAGEDy correctly aligned these datasets: the WT and the KO datasets initially align with one another, but the KO finishes its full process before the WT. For comparison, we also applied cellAlign and Genes2Genes (G2G), another trajectory alignment method which mainly focuses on aligning single genes, but can also be used to align multiple genes at once (Sumanaweera et al. 2023). Like TrAGEDy, the cellAlign distance matrix revealed the initial alignment of the WT and KO, before the KO trajectory ends while the WT continues. However, cellAlign failed to position the correct end point of the KO trajectory relative to the WT due to running DTW on the whole dissimilarity matrix. G2G also returns an incorrect alignment, suggesting that the simulated WT and KO datasets have a 1:1 alignment, although the number of matched genes does decrease as the trajectories progress (Fig.9B).

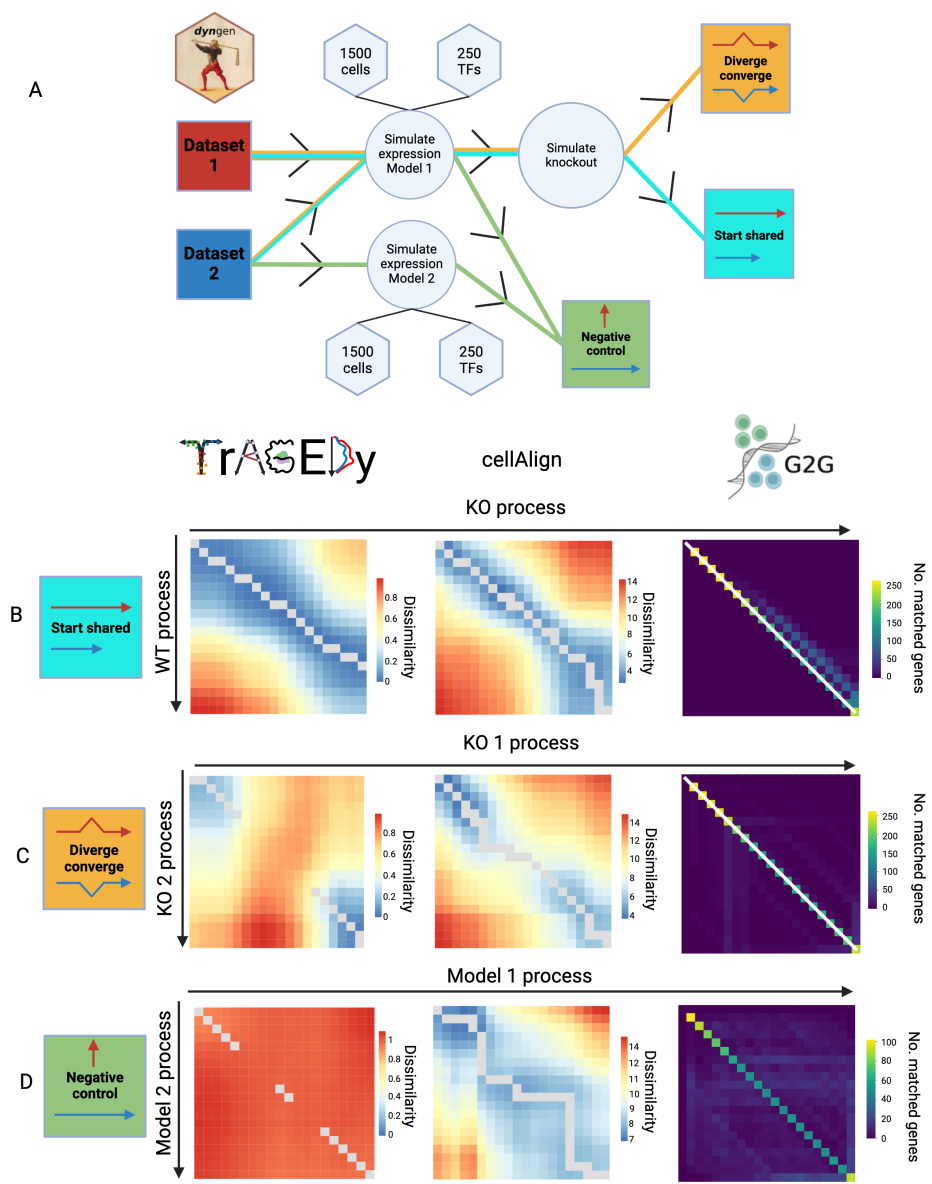


Figure 9: Alignment of simulated single cell RNA sequencing datasets with TrAGEDy and cellAlign.

Schematic showing the process of generating the three different topologies of alignment datasets, each with 1,500 cells and 250 transcription factors which drive the simulated process. For the diverge converge and start shared datasets, two datasets were generated using the same simulated expression model before knockouts were simulated for these datasets. For the negative control dataset, the expression of the two datasets was simulated using two different models (A). Heatmaps showing the alignment of the start shared (B), diverge converge (C) and negative control (D) Dyngen simulated datasets as assessed by TrAGEDy, cellAlign and Genes2Genes (G2G). Each box on the heatmap represents the alignment score of an interpolated point of each of the two datasets. For TrAGEDy and cellAlign the score represents transcriptional dissimilarity (as assessed by Spearman's correlation and Euclidean distance respectively), while G2G shows the number of genes whose expression patterns is matched between the two interpolated points. The grey line represents the path of optimal alignment for TrAGEDy and cellAlign, while in G2G the optimal alignment is represented by a white line. The alignment heatmap of G2G for the negative control dataset has been edited to remove the alignment line; reflecting its conclusion that the two datasets share no common alignment.

To assess how the methods handle a scenario where the trajectories have shared start and end points, but deviate in the middle, we simulated two datasets with Dyngen that have a diverge-converge backbone. By knocking out one of the divergent branches in one dataset (KO 1) and the other divergent branch in the other dataset (KO 2), we simulated linear trajectories that do not share a common process in their middle sections (Fig.9A). Applying TrAGEDy to this scenario accurately captured the initial and terminal alignment of the two trajectories, while leaving the middle sections unaligned. cellAlign failed to align KO 1 and KO 2 correctly, as cellAlign does not include functionality to prune matches. G2G also returned an incorrect alignment, as it suggested that there was a 1:1 alignment between the two simulated KO datasets (Fig.9C) despite the fact that they two datasets are transcriptionally divergent in the middle of their processes.

For the negative control, two datasets with distinct gene regulatory networks and transcription kinetics models were generated (referred to as model 1 and model 2) (Fig.9A). As there is no shared process between the conditions, the dissimilarity scores are expected to be high, and no alignment should be found. The dissimilarity score calculated by TrAGEDy between model 1 and model 2 showed that most of the matches have a dissimilarity score close to 1, and thus do not share a common process. As TrAGEDy finds the optimal path by looking at the context of the whole dissimilarity matrix, even if the dissimilarity scores are high, it will still find a path, as seen in the negative control (Fig.9C). TrAGEDy will give a warning to the user if the median of the cut dissimilarity matrix scores is over 0.6 and, thus the user must decide how to interpret the resulting alignment in the context of the overall dissimilarity score of the path. CellAlign also creates a path through the dissimilarity matrix, which has a higher overall dissimilarity range than the other two datasets, with the Euclidean distance ranging from $\sim 7-14$ (Fig.9D), while the other two datasets ranged from $\sim 4-14$ (Fig.9B, C). If, instead, Pearson correlation is selected for cellAlign, it becomes clearer from the alignment heatmap that the two datasets share little transcriptional similarity (appendix Fig.5C). As G2G directly models insertion and deletions in its alignment algorithm it can determine that there is no common alignment between the two datasets and, thus, is the only method to show that there is no underlying alignment in its alignment path (Fig.9D).

These analyses show that TrAGEDy is capable of faithfully capturing the underlying alignments of simulated biological scRNA-seq datasets, where current methods fail.

9.2 WT vs *ZC3H20* KO *T. brucei* alignment

Simulated data represents a ‘best case’ scenario for benchmarking tools, and often lacks the complexity of real datasets. We thus decided to apply TrAGEDy to real scRNA-seq datasets where the underlying processes have been characterised. For our first application, we examined a distinct developmental process in the kinetoplastid parasite *T. brucei*. In the bloodstream of its mammalian host, WT *T. brucei* parasites transition from replicative slender forms to non-replicative stumpy forms, through a quorum sensing mechanism (Reuner et al. 1997, Dean et al. 2009, Rojas et al. 2019 & Matthews 2021). These stumpy forms are preadapted to survive within the tsetse fly, which acts a vector for parasite transmission between mammals (Rico, Rojas, et al. 2013). Mutant *T. brucei* where an RNA binding protein essential for slender to stumpy differentiation, *ZC3H20*, has been knocked out (*ZC3H20* KO) are unable to undergo this transition (B. Liu, Kamanyi Marucha, and Clayton 2020, Cayla et al. 2020) and fail to express stumpy associated genes. Briggs and colleagues (Briggs, Rojas, et al. 2021) used 10X Chromium to sequence two biological replicates of WT *T. brucei* parasites (WT01 and WT02) undergoing this transition *in vitro*, as well as *ZC3H20* KO failing to differentiate in the same conditions. The authors defined four main clusters, Long Slender (LS) A and B, and Short Stumpy (SS) A and B, to show the progression of the transition, with LS A being the start and SS B being the end. Thus, the final alignment is expected to show that the WT and the *ZC3H20* KO datasets share an initial common developmental process, but that the KO process is truncated or branched relative to the WT, as suggested by previous

analysis using data integration and Slingshot TI.

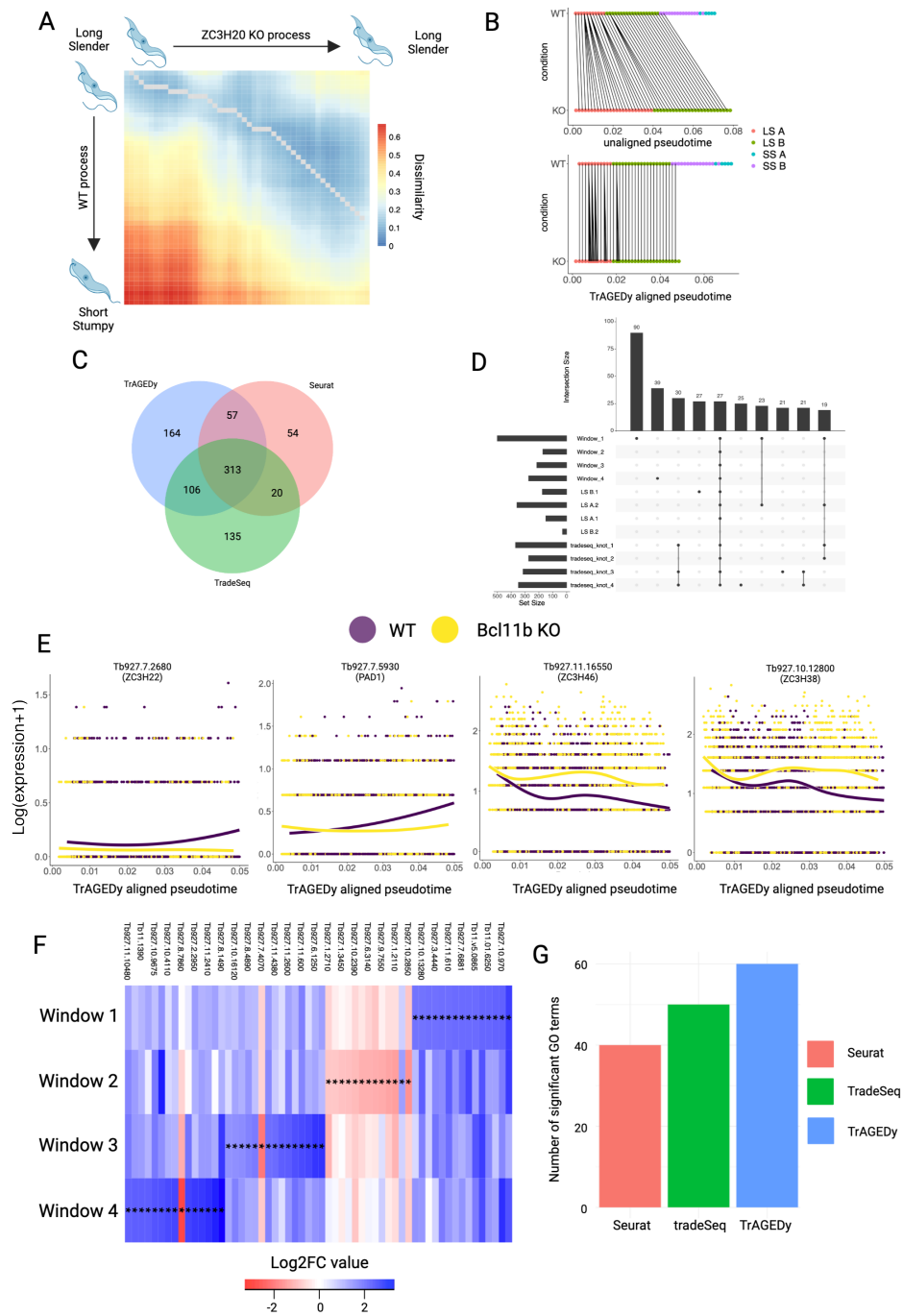


Figure 10: Analysis of WT and *ZC3H20* KO *T. brucei* development with TrAGEDy.

TrAGEDy alignment of the WT and *ZC3H20* KO trajectories of *T. brucei* development with transcriptional dissimilarity calculated with Spearman correlation (A). TrAGEDy alignment of the interpolated points of the WT and *ZC3H20* KO trajectories, showing connections between the two processes. The interpolated points are displayed across pseudotime before and after the pseudotime values have been modified by TrAGEDy (B). Venn diagram showing the intersection of the DE genes captured by TrAGEDy, TradeSeq and Seurat (C). UpSet plot showing the top 10 intersections of DE genes captured by TrAGEDy across 4 windows of comparison and Seurat over 4 clusters and TradeSeq over 4 knot comparisons (D). TradeSeq plotSmoother plots showing the expression changes that occur across the WT (yellow) and *ZC3H20* KO (purple) trajectories within the cells of aligned pseudotime, as calculated by TrAGEDy, which share a common process. The 4 genes plotted were only identified as being significantly DE by TrAGEDy (E). Heatmap showing DE genes with the highest log 2 fold-change (Log_2FC) captured (but not uniquely captured) between the WT and *ZC3H20* KO trajectories at 6 different windows of comparison. Each index is coloured by the Log_2FC (red = higher expression in KO, blue = higher expression in WT) and a * symbol indicates the difference is significant (Bonferroni corrected p-value < 0.05, Mann-Whitney U test, $\text{abs}(\text{Log}_2\text{FC}) > 0.5$, minimum percentage expressed > 0.1). A full list of heatmap genes can be found in supplementary table 7 (F). Barplot showing the number of significant Gene Ontology (GO) terms (Benjamini-Hochberg adjusted p-value < 0.01) returned when GO enrichment analysis was carried out using TriTrypDB on all the DE genes returned by TrAGEDy, TradeSeq and Seurat.

PHATE (Moon et al. 2019) embeddings were generated independently for the three datasets: two WT replicates (WT01 and WT02) and the *ZC3H20* KO. The PHATE embeddings were used as the basis for Slingshot TI to calculate pseudotime values for each of the cells. Applying TrAGEDy to WT01 and WT02 returned an alignment that started at the same point, and ended just before the final endpoint, returning a nearly 1:1 alignment, as predicted of biological replicates (appendix Fig.6). We next used the aligned WT values to perform TrAGEDy alignment between the combined WT replicates and the *ZC3H20* KO dataset. The resulting alignment showed that the two process have an initial point of alignment but the *ZC3H20* KO process was truncated relative to the WT (Fig.10A). Using the DTW calculated path, we then adjusted the pseudotime values of the trajectories to get a common pseudotime axis (Fig.10B).

By aligning trajectories rather than integrating the datasets, we expected to preserve more variability between them, which may be reflected in the DE genes captured. We therefore compared TrAGEDy against other methods that can extract DE genes between conditions. We compared the differential gene expression tests by TrAGEDy with established methods of extracting DE genes: we compared TrAGEDy to a standard Seurat V5 pipeline (Y. Hao, T. Stuart, et al. 2024) of comparing DE genes between conditions, and TradeSeq’s method of comparing DE genes across trajectories from different conditions. DE genes were defined as genes whose absolute log 2 fold change (Log_2FC) was more than 0.75, Bonferroni corrected p-value was less than 0.05, and were expressed in at least 10% of the cells in one of the comparison groups. To ensure comparability between the methods, the Log_2FC of a gene between conditions was calculated using the same formula across all the different methods, namely the formula used in Seurat V5.

TrAGEDy detected more significant DE genes than both Seurat and TradeSeq (Fig.10C, D): TrAGEDy uniquely captured 164 genes as being DE, while TradeSeq and Seurat only returned 135 and 54, respectively (Fig. 10C, appendix file 1, 2, 3). Furthermore, the genes captured only by TrAGEDy that were more highly expressed in the WT were consistent with the slender to stumpy transition. These included the Arginine Kinase *AK3* (Tb927.9.6210) (Ooi et al. 2015), RNA binding proteins *ZC3H22* and *RBP38* (Tb927.7.2680, Tb927.11.5850) (Erben et al. 2021, Sbicego et al. 2003), Protein Associated with Differentiation 1 (*PAD1* - Tb927.7.5930) (Dean et al. 2009) and delta-1-pyrroline-5-carboxylate dehydrogenase

(*DP5CDH*) (Tb927.10.3210) (Fig.10F, appendix file 1). For *ZC3H20* KO associated genes, two zinc finger genes, *ZC3H38*, *ZC3H46* (Tb927.10.12800, Tb927.11.16550) (Lueong et al. 2016), were found to be DE towards the end of the shared process. Plotting the smoothed expression of some of these DE genes unique to TrAGEDy showed that the expression patterns matches the TrAGEDy output (Fig.10E). In terms of runtime required to return DE results, both TrAGEDy and Seurat returned results in under a minute, while TradeSeq took around 2 hours (appendix file 4).

Returning more significant DE genes and attaching biological meaning to individual genes are not robust metrics for assessing how well a DE gene detection method is performing, as they are prone to the influence of bias. As an example, randomly assigning 75% of the genes in the dataset to be DE would end up with more DE genes returned than most DE gene detection methods. Furthermore, there is no direct ground truth for whether the genes captured are DE or not. To address these issues, we used the number of significant GO terms returned from each methods DE gene list as a proxy for how biologically meaningful the output of the DE gene detection methods were likely to be. This approach was previously utilised in a previous analysis (Song and J. J. Li 2021), to test the usefulness of different DE gene detection methods (Song and J. J. Li 2021). We ran biological process GO term analysis on the three methods' lists of DE genes, with GO term significance being set at Benjamini Hochberg procedure adjusted p-value < 0.01. Looking at the GO terms generated from the DE gene lists of the three methods showed that TrAGEDy returned 60 significant GO terms compared to 50 from TradeSeq and 40 from Seurat (Fig.10G) (appendix file 5), with all but three of the GO terms returned by Seurat also returned by TrAGEDy and only 15 GO terms being unique to TradeSeq (appendix Fig.7B). When GO terms were generated on the list of uniquely captured genes for each method, TrAGEDy was the only method to return any significant GO terms (appendix Fig.7A, appendix file 6).

The above results show that TrAGEDy reveals more biologically relevant information on the slender to stumpy transitions under WT and *ZC3H20* KO conditions than established protocols.

9.3 WT vs *Bcl11b* KO T cell development analysis

Through applying TrAGEDy to the *T. brucei* datasets, we replicated an analysis that has already been carried out. With our next application, we used TrAGEDy in an analysis scenario which has not been attempted. For this we applied TrAGEDy to a scRNA-seq dataset of *in vitro* T cell development in WT and *Bcl11b* KO mice over two timepoints (W. Zhou et al. 2022). The data contain cells undergoing the early stages of T cell development in the thymus, from thymus-seeding precursors (TSP) to around the Double negative (DN) 4/Double positive stage. The authors reported that when *Bcl11b* expression is knocked out, the cells deviate from the T cell commitment pathway and develop distinct transcriptomic signatures around the DN2 stage of T cell development. However, some *Bcl11b* KO cells still express markers of T cell pathway commitment, such as the gene for the pre-T cell Antigen receptor α (*Ptcra*) (Hwang et al. 2020) and *Rag1* (appendix Fig.8), one of the central genes responsible for T cell receptor gene recombination, suggesting there may be some KO cells that have reached the post DN2 stages where the cells express a pre-TCR receptor and undergo β -selection (Mombaerts et al. 1992). In this context, we aimed to use TrAGEDy to compare the WT and *Bcl11b* KO trajectories of T cell development and attempted to identify differences in transcriptomes that occur between the WT and *Bcl11b* KO as they continue down the normal path of T cell development.

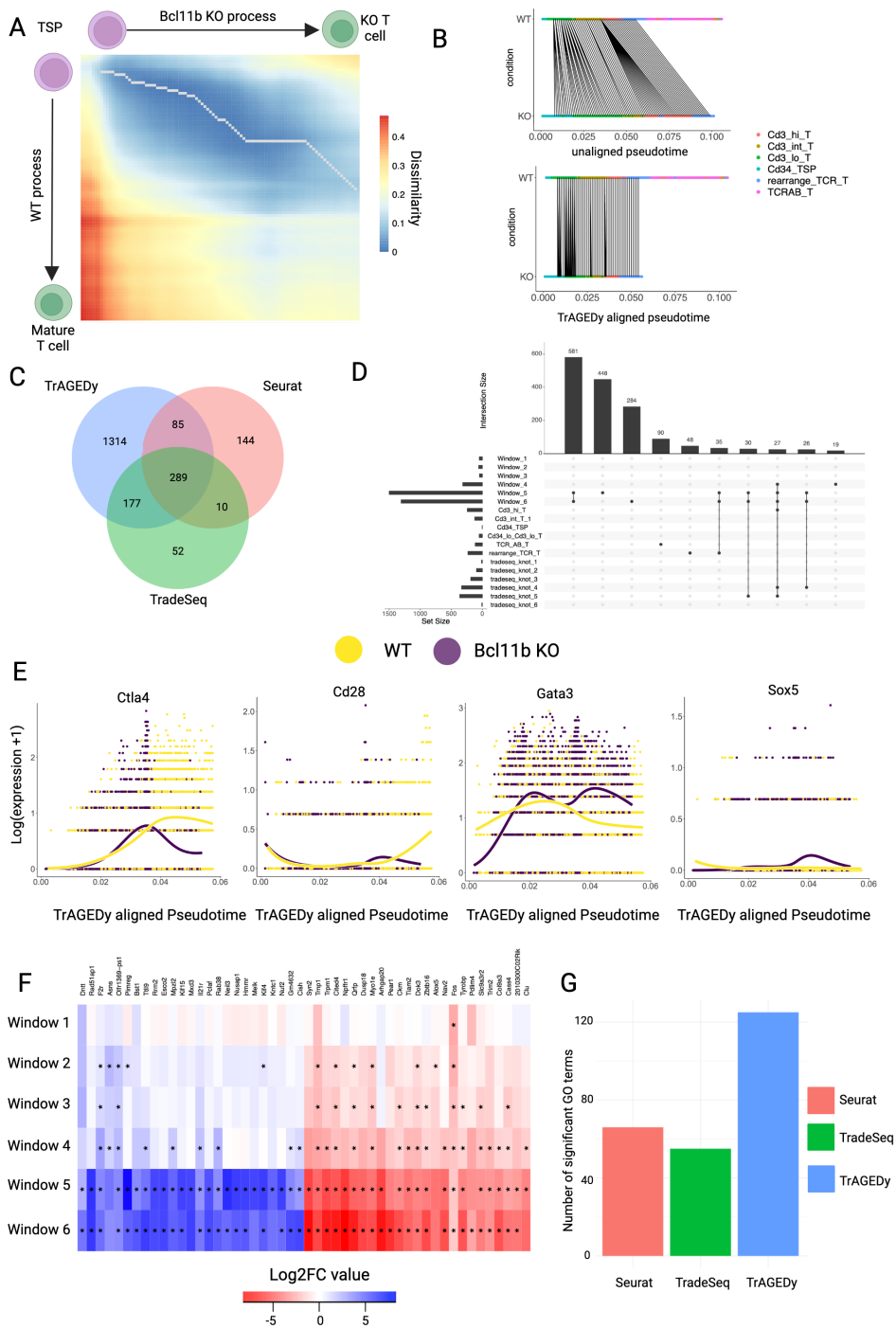


Figure 11: Analysis of WT and *Bcl11b* KO trajectories of T cell development with TrAGEDy.

TrAGEDy alignment of the WT and *Bcl11b* KO trajectories of T cell development with transcriptional dissimilarity calculated with Spearman correlation (A). TrAGEDy alignment of the interpolated points of the WT and *Bcl11b* KO trajectories, showing connections between the two processes. The interpolated points are displayed across pseudotime before and after the pseudotime values have been modified by TrAGEDy (B). Venn diagram showing the intersection of the DE genes captured by TrAGEDy, TradeSeq and Seurat (C). UpSet plot showing the top 10 intersections of DE genes captured by TrAGEDy across 6 windows of comparison and Seurat over 6 clusters and TradeSeq over 6 knot comparisons (D). TradeSeq plotSmoother plots showing the expression changes that occur across the WT (yellow) and *Bcl11b* KO (purple) trajectories within the cells of aligned pseudotime, as calculated by TrAGEDy, which share a common process. The 4 genes plotted were only identified as being significantly DE by TrAGEDy (E). Heatmap showing DE genes captured (but not uniquely captured) by TrAGEDy with the highest log₂ fold-change (Log₂FC) between the WT and *Bcl11b* KO trajectories at 6 different windows of comparison. Each index is coloured by the Log₂FC (red = higher expression in KO, blue = higher expression in WT) and a * symbol indicates the difference is significant (Bonferroni corrected p-value < 0.05, Mann-Whitney U test, abs(Log₂FC) > 0.5, minimum percentage expressed > 0.1). Full list of genes can be found in appendix file 13(F). Barplot showing the number of significant Gene Ontology (GO) terms (Benjamini-Hochberg adjusted p-value < 0.01) returned when GO enrichment analysis was carried out using clusterProfiler on all the DE genes returned by TrAGEDy, TradeSeq and Seurat.

The cells were sequenced over two runs, with each run containing a variety of biological and technical replicates, as well as cells from different sampling days. For simplicity, the WT cells sequenced in run 1 will be referred to as WT1, the *Bcl11b* KO cells sequenced in run 1 will be referred to as *Bcl11b* KO1, and those sequenced in run 2 as WT2 and *Bcl11b* KO2, resulting in four individual datasets. When projected into the PHATE space, some of the WT1 cells were separated from the main body of the trajectory, which had downstream effects on the alignment, and so these cells were removed because TrAGEDy cannot function when there are gaps in the pseudotime axis (appendix Fig.9). TrAGEDy was then used to first align the two datasets for each condition together, resulting in two separate alignments: a WT alignment of WT1 and WT2, and a *Bcl11b* KO alignment of *Bcl11b* KO1 and *Bcl11b* KO2 (appendix Fig.10). The WT1 dataset only contains cells from day 10 of sampling, while WT2 contains cells from day 10 and 13. TrAGEDy captures this difference, showing a strong initial alignment of the two WT datasets with the WT1 dataset finishing before the WT2 (appendix Fig.10A). TrAGEDy was then performed on the aligned WT and *Bcl11b* KO trajectories. The resulting final alignment shows the two conditions sharing an initial common process, but the *Bcl11b* KO developmental progression finishes around the point when cells transition to express both TCR α and β chain genes (Fig.11A, B).

Overall, TrAGEDy returned 1865 DE genes across 6 windows of comparison, while TradeSeq and Seurat captured 528 DE genes across 6 knots and 6 cluster comparisons, respectively (Fig.11C, appendix file 8, 9 & 10). All the methods found more DE genes towards the end of the biological process compared with the beginning (Fig.11D). For genes only captured by one of the methods, TrAGEDy uniquely captured 1314 DE genes, while Seurat returned 144 and TradeSeq 52 (Fig. 11C, appendix file 8, 9 & 10). Genes with higher expression in the WT that only TrAGEDy identified as being DE include those associated with T cell signaling (*Cd28*, *Ctla4*) (Sansom 2000), cell cycle associated genes (*Top2a*, *Dut*, *Cenpa*) (Giotti et al. 2019), and those with functions for protecting against reactive oxygen species (ROS) (*Prdx4*, *Prdx1*) (Rhee et al. 2012). The genes only TrAGEDy identified as being significantly higher expressed in *Bcl11b* KO cells include pro-apoptotic factors (*Ikbip*, *G0s2* and *Stk4*) (Hofer-Warbinek et al. 2004, Welch et al. 2009, Cinar et al. 2007) and transcription factors (*Gata3* and *Batf*) (Wan 2014, Betz et al. 2010) (appendix file 8). Plotting some of the DE genes that only TrAGEDy captured with TradeSeq smoothed expression show they are concurrent with the patterns of expression TrAGEDy suggests, with *Cd28* and

Ctla4 being more highly expressed in WT T cells, while *Gata3* and *Sox5* are more highly expressed in *Bcl11b* KO cells (Fig.11E). The times taken to return DE results was longer than the *T. brucei* analysis for all methods, however Seurat still managed to return results within seconds and TrAGEDy only took just over two minutes. In contrast, TradeSeq took over 20 hours to return a result (appendix file 4).

TrAGEDy found the most significant GO terms compared to Seurat and TradeSeq (Fig.11G, appendix file 11). Furthermore, when only the uniquely captured DE genes were utilised, TrAGEDy was the only method that returned any significant GO terms (appendix fig.11A, appendix file 12). Of the significant GO terms returned by the methods for their entire DE gene list, only 14 were shared between all three methods (appendix fig.11B), with many of them being GO terms related to regulation of lymphocyte development and activation (appendix file 11). Of the GO terms that only TrAGEDy captured, most of them were related to regulation and activation of cell cycle-associated processes (appendix file 11 & 12).

10 Chapter 1: Materials and Methods

10.1 The TrAGEDy method

The following is an overview of the steps that TrAGEDy takes to analyse a dataset. A workflow diagram of the TrAGEDy process and its core steps can be seen in appendix figure 1.

10.1.1 Create interpolated points

Performing alignment directly on every cell in a dataset would be computationally and time expensive and prone to noise. As such, a modification of the cellAlign method was used to smooth gene expression over the trajectory by creating a user-defined number of interpolated points that sample the gene expression patterns of surrounding cells at specific points in the process and perform alignment on them. Each interpolated point is given the same sized set window of pseudotime around it. Cells within that window contribute highly to the gene expression of the interpolated point, while ones further away contribute less. TrAGEDy, allows the user to specify a window size, but this value is then weighted by the density of cells around each interpolated point. Interpolated points with many cells around it will have a smaller window, while those with few surrounding cells will have a larger window (appendix Fig.1B).

10.1.2 Scoring dissimilarity

For every interpolated point in the two trajectories, we calculate the transcriptomic dissimilarity between it and all the interpolated points on the other trajectory and store it in a matrix. This can be done using any metric, e.g. Pearson correlation, Spearman correlation, Euclidean distance, given the right processing of the data (e.g. scaling prior of gene expression prior to Euclidean distance). For our experiments, Spearman correlation was used to assess dissimilarity. We change the correlation scores to be 1 minus the correlation score hence, all the dissimilarity scales start from 0 (i.e. least dissimilarity) (appendix Fig.1C).

10.1.3 Identifying the optimal path of the two trajectories

To find the optimal alignment between the interpolated points of the two trajectories, Dynamic Time Warping (DTW) was utilised. DTW requires that every point on the two processes is matched with

at least one other point on the opposing process. It also requires that the first and last points of each process must be matched to one another. To tackle this constraint, TrAGEDy scans the first and last row and column of the dissimilarity matrix and identifies the indexes with the lowest score from the beginning and end. These become the initial start and ends points of the process. To make sure no valid matches are missed from the downstream analysis, TrAGEDy takes the following steps. First, TrAGEDy finds out how much the absolute dissimilarity changes across the row/column of the beginning and end changes as we move across them (appendix Fig.2A). The user then defines whether to find the median or mean of these points and this value is used as a cutting threshold. To give a value to each index to compare against the cutting threshold, TrAGEDy finds the absolute difference between the scores of each index in the chosen row/column and the current start/end point (appendix Fig.2B). If there is an index value which occurs before the current start point or after the current end point whose dissimilarity score is less than the threshold, it is considered as a possible start/end point (appendix Fig.2B). For each possible path, we remove the interpolated points that fall before the start point and after the end point before performing DTW (appendix Fig.2C). To see what paths give us the best dissimilarity scores, we bootstrap the dissimilarity scores of the matched interpolated points along the path (appendix Fig.2C). The user can set the number of iterations during bootstrapping but the number of samples taken during each iteration set as the length of the longest path. The mean bootstrapped dissimilarity scores for each path is then calculated.

To deal with the scenario where two processes might differ somewhere in the middle of their processes, TrAGEDy first finds two thresholds for the matched and unmatched indexes (appendix Fig.3A). The user chooses to use either the mean or median of the dissimilarity score of all the unmatched indexes and all the matched indexes. If a matched index's dissimilarity score is closer to that of the unmatched threshold than the matched threshold then it is cut, otherwise it remains matched (appendix Fig.3B).

10.1.4 Align pseudotime of interpolated points and cells

The aim of aligning the pseudotime of the interpolated points is to give matched interpolated points similar pseudotime values (appendix Fig.1D). As we have two processes, we have two vectors:

$$y_i^1 \{i \in 1 : m\}$$

$$y_j^2 \{j \in 1 : n\}$$

where m and n are the number of interpolated points on process 1 and 2 respectively, with y_i^1 being the pseudotime of the i th interpolated point in process 1 and y_j^2 being the pseudotime of the j th interpolated point in process 2.

Interpolated points may only match one another or be a part of a multi-match, where one interpolated point on one of the trajectories is connected to 2 or more interpolated points on the other trajectory. For adjusting pseudotime of the former scenario (appendix Fig.4A, Equation 1), given that y_i^1 and y_j^2 are matched and that $y_j^2 < y_i^1$:

$$y_{i:m}^1 = y_i^1 - i : m + (y_j^2 - y_i^1)$$

For a multi-match, first TrAGEDy aligns the pseudotime of the first section of the multimatch (appendix Fig.4B, equation 2), using the steps outlined for the individual match, then it does the same

for the next match which isn't in the multi-match (appendix Fig.4C, equation 3). The interpolated points that are multi-matched to one interpolated point on the other process, have their pseudotime values scaled between the aligned pseudotime value of the first match in the multi-match and the aligned pseudotime value of the match that occurs after the multi-match, adjusted by the difference between pseudotime of the next match and the pseudotime of the first match in the multi-match, normalized by the number of interpolated points being scaled (appendix Fig.4D, equation 4). This leaves us with an aligned pseudotime axis, where the matched interpolated points have similar pseudotime values (appendix Fig. 4E). We used the cellAlign method of mapping gene expression values from individual cells onto interpolated points to map pseudotime values from interpolated points onto the individual cells.

10.1.5 Differential expression analysis with TrAGEDy

TrAGEDy identifies DE genes across pseudotime, between the two conditions by taking a sliding window soft clustering approach (appendix Fig.1E).

First the cells were assigned to the closest interpolated point in terms of pseudotime via one round k-means clustering, with the interpolated points acting as the cluster centroid. The interpolated points that were matched together were grouped, with the cells that were assigned to these interpolated points acting as a cluster which the window slides over when making comparisons. The user defines how many windows of comparison will be made across the pseudotime, and how much overlap there will be, in terms of matched interpolated points, between each successive window.

The user chooses a method of assessing significance using either a t-test or a Mann-Whitney U test. Log₂ fold change (Log₂FC) of the genes between conditions is calculated using the same formula used by Seurat V5. To adjust the p-values to account for multiple testing, Bonferroni correction is used across all the genes in the dataset.

10.2 Simulated dataset creation and benchmarking

10.2.1 Simulating trajectories with Dyngen

All simulated datasets were simulated with Dyngen (Cannoodt et al. 2021). The first positive control experiment trajectories were simulated with a linear backbone with parameters drawn from the same transcription factor, feature and kinetic network distributions. One of the datasets was not modified beyond the standard pipeline, while the other had a full knockout in the B5 gene module, causing a stunted process in comparison to the other dataset. The second positive control experiment trajectories were simulated with a bifurcating-converging backbone with parameters drawn from the same transcription factor, feature and kinetic network distributions. One of the datasets had a full knockout in the C1 gene module while the other dataset had a full knockout in the D1 gene module. The negative control experiment trajectories were simulated with a linear backbone with parameters drawn from different transcription factor, feature and kinetic network distributions. The specific parameters and program calls used to simulate the individual datasets can be found on the TrAGEDy (chapter 1) GitHub link in the "Code Availability" section.

10.2.2 Applying TrAGEDy, Genes2Genes and cellAlign to simulated trajectories

All simulated datasets were processed using the following. In Seurat, the simulated counts were normalized prior to scaling across all simulated genes. Principal Component Analysis (PCA) was performed on the scaled data and using the first 5 PCA dimensions clustering was performed at a resolution of 0.2 for all the datasets as well as UMAP. Slingshot (Street et al. 2018) was performed to get pseudotime values for each of the cells, using the UMAP space as the basis for the trajectory. To get the feature space, the top 100 DE genes in terms of Log_2FC for each cluster in each dataset were collected for both datasets being compared. The prior analysis steps were kept the same for both cellAlign and TrAGEDy analysis, as well as the window size and the number of interpolated points. For cellAlign Euclidean distance and Pearson correlation was used to calculate dissimilarity, while for TrAGEDy Spearman correlation was used. For Genes2Genes, a binning number of 20 was used for all the comparisons and the same feature space was used as in the TrAGEDy and cellAlign analyses. More specific parameter values used for each dataset can be found on the TrAGEDy (chapter 1) GitHub, linked in “Code Availability” section.

10.3 Analysis of WT vs *ZC3H20* KO *T. brucei* scRNA-seq dataset

10.3.1 Pre-processing of WT vs *ZC3H20* KO *T. brucei* dataset

The processed R dataset (RDS) file containing the final cluster labels was acquired from the authors. The ‘LS A.1’ and ‘LS A.2’ clusters were merged into the cluster ‘LS A’ and the ‘LS B.1’ and ‘LS B.2’ were merged into the cluster ‘LS B’.

10.3.2 TrAGEDy analysis of WT vs *ZC3H20* KO *T. brucei* dataset

The integrated object was split into the individual datasets (WT01, WT02 and *ZC3H20*_KO). The top 200 DE genes (defined as absolute $\text{Log}_2\text{FC} > 0.75$, Bonferroni corrected p-value < 0.05 and the gene expressed in at least 25% of cells) across the clusters were found for each dataset and used as the basis for constructing the PHATE space and assessing similarity with TrAGEDy. The number of genes in the feature space amounted to 455 genes. PHATE embeddings were constructed from the normalised gene expression count matrix. Pseudotime was calculated using Slingshot for each of the three datasets, with the ‘LS A’ cluster used as the starting point. The WT01 and WT02 datasets were then aligned with TrAGEDy, with 50 interpolated points used to align the processes. This TrAGEDy aligned WT dataset was then aligned against the *ZC3H20* KO dataset. Differential expression was carried out over 4 windows, with a gene identified as being DE if the absolute Log_2FC was more than 0.75, Bonferroni corrected p-value < 0.05 and the gene expressed in at least 10% of cells in one of the conditions.

10.3.3 Seurat analysis of WT vs *ZC3H20* KO *T. brucei* dataset

The ‘SS A’ and ‘SS B’ clusters were removed from the integrated dataset, as they are not present in the *ZC3H20* KO dataset, and the FindMarkers function was performed for all the four slender clusters (LS A.1, LS A.2, LS B.1 & LS B.2) between the WT and *ZC3H20* KO conditions. A gene identified as being DE if the absolute Log_2FC was more than 0.75, Bonferroni corrected p-value < 0.05 and the gene expressed in at least 10% of cells in one of the conditions.

10.3.4 TradeSeq analysis of WT vs *ZC3H20* KO *T. brucei* dataset

The ‘SS A’ and ‘SS B’ clusters were removed from the integrated dataset. Using the authors PHATE space, Slingshot was applied to build trajectories. The optimal number of knots for fitting the general additive model to the dataset was assessed through the evaluateK function, with the optimal number determined to be 8. fitGAM was run on the dataset followed by the conditionTest function to identify DE genes between the WT and *Bcl11b* KO conditions. 4 knot comparisons were made, and a gene were determined to be DE if the absolute Log_2FC was greater than 0.75, Bonferroni corrected p-value < 0.05 and the gene was expressed in at least 10% of the cells that fell within the current knot comparison, in one of the conditions.

10.4 GO term analysis of *T. brucei* DE genes

Biological process GO term analysis of the lists of DE genes (either unique DE genes or all of them) for the three methods was carried out using TriTrypDB. GO terms whose Benjamini-Hochberg corrected p-value was less than 0.01 were identified as significantly enriched GO terms.

For all cases, genes were analysed in TriTrypDB to get their gene short names and functions (Amos et al. 2022 & Alvarez-Jarreta et al. 2024).

10.5 Analysis of WT vs *Bcl11b* KO *in vitro* T cell development scRNA-seq dataset

10.5.1 Preprocessing of WT vs *Bcl11b* KO *in vitro* T cell development dataset prior to TrAGEDy analysis

The RDS file containing the scRNA-seq information was acquired from the authors (W. Zhou et al. 2022).

The dataset was split into the WT and *Bcl11b* KO conditions and then further split into the different sequencing runs. For each of the four scRNA-seq objects derived from the previous, the data was analysed using the Seurat V5 pipeline. When the data was scaled, the effects of the cell cycle were regressed out by passing the cell cycle associated genes Seurat provides into the var.to.regress parameter. For each dataset, PCA was carried out on the scaled-regressed expression matrix, reducing the dimensions down to 50. An elbow plot was used to determine the number of Principal Components (PC) used to create a UMAP space embedding. Clustering was performed using the same number of PCs for UMAP embedding calculation. Clustree (Zappia and Oshlack 2018) was used to help choose a resolution for clustering that leads to stable clusters. Clusters were annotated as follows. Clusters that did not express *Cd3* genes but expressed *Cd34* were defined as ‘Cd34_TSP’, clusters which did not express *Cd34* or *Cd3* were characterized as ‘early_T’, clusters which had low expression of *Cd3* but no *Cd34* expression were characterised as ‘Cd34_lo_Cd3_lo_T’, those with high *Cd3* expression were ‘Cd3_T_hi’ and those with middle levels of *Cd3* expression were “Cd3_T_mid”. Clusters which had high expression of *Cd3* and expressed *Ptcra* and *Rag1* were defined as ‘rearrange_TCR_T’ and those that expressed the TCR signal transduction molecule *Zap70* and some cells that express the gene *Trac*, were defined as ‘TCR_AB_T’. Clusters that express *Rora* were defined as ‘Rora_T’, clusters that expressed high levels of interferon associated genes were defined as ‘Interferon_response’ and cells which clustered out with the main body of the UMAP were classed as ‘Outlier’. The ‘Interferon_response’ and “Outlier” clusters were removed before downstream analysis with TI, TrAGEDy, Seurat DE or TradeSeq.

10.5.2 TrAGEDy analysis of WT vs *Bcl11b* KO *in vitro* T cell development dataset

In order to build a gene feature space to construct PHATE embeddings and assess dissimilarity with TrAGEDy, the feature space for the dataset was made of the combined DE genes for each cluster whose absolute $\text{Log}_2\text{FC} > 0.5$, Bonferroni corrected p-value < 0.05 and was expressed in at least 1%, across all four of the datasets. To reduce the effect of the cell cycle, genes supplied by Seurat as being important in the cell cycle and genes included in the Mouse Genome Informatics GO term ‘cell cycle’ were removed from the feature space. For all four of the individual datasets, 10 dimensional PHATE embeddings were generated from the normalised gene expression count matrix for the features selected previously. For the WT1 dataset, cells with a pseudotime higher than 0.1 (11 cells) were removed. TI was then carried out using Slingshot on the PHATE embeddings with the Cd34_TSP cluster chosen as the starting point. TrAGEDy was carried out on each of the conditions, using 100 interpolated points with a window size of the maximum pseudotime value, from either of the two datasets being analysed, divided by 90. TrAGEDy aligned pseudotime values were then created, resulting in a TrAGEDy aligned WT and a TrAGEDy aligned *Bcl11b* KO dataset.

The WT and *Bcl11b* KO TrAGEDy aligned datasets were then analysed using TrAGEDy. TrAGEDy was carried out with the same number of interpolated points and window size as the replicate alignments. TrAGEDy differential expression test was then carried out across 6 windows of comparison. A gene was returned as being DE if the absolute Log_2FC was greater than 0.75, Bonferroni corrected p-value < 0.05 and the gene was expressed in at least 10% of cells in the window of comparison, in one of the conditions. The final TrAGEDy aligned WT and *Bcl11b* KO datasets were then analysed using TradeSeq to create smooth gene expression plots across the aligned pseudotime.

10.5.3 Seurat analysis of WT vs *Bcl11b* KO *in vitro* T cell development dataset

Using Seurat V5 integration (Satija et al. 2015, A. Butler et al. 2018, T. Stuart et al. 2019 & Y. Hao, S. Hao, et al. 2021, Y. Hao, T. Stuart, et al. 2024), the four datasets were integrated using the “integrated.cca” method. When the data was scaled, the effects of the cell cycle were regressed out by passing the cell cycle associated genes Seurat provides into the var.to.regress parameter. PCA was carried out on the scaled-regressed expression matrix, reducing the dimensions down to 50. An elbow plot was used to determine the number of PCs used to create a UMAP space embedding. Clustering was performed using the same number of PCs for UMAP embedding calculation. Clustree was used to help choose a resolution for clustering that leads to stable clusters and clusters were annotated as described previously. The FindMarkers function was used to identify DE genes between the conditions for each cluster. Genes were determined to be DE if the absolute Log_2FC was greater than 0.75, Bonferroni corrected p-value < 0.05 and the gene was expressed in at least 10% of the cells in one of the conditions for the cluster being compared.

10.5.4 TradeSeq analysis of WT vs *Bcl11b* KO *in vitro* T cell development dataset

Using the Seurat integrated, cell cycle regressed dataset; PHATE was carried out, using the scaled and cell cycle regressed gene expression matrix. Slingshot was then applied to the dataset, using the Cd34_TSP cluster as the starting point of the trajectory. Cells on the trajectory path which ended with the TCR_AB_T cluster was kept and all other cells were removed from the dataset. The optimal number of knots for fitting the general additive model to the dataset was assessed through the evaluateK function, with the optimal number determined to be 7. fitGAM was run on the dataset followed by the conditionTest function to identify DE genes between the WT and *Bcl11b* KO conditions. 6 different

knot comparisons were made, and a gene were determined to be DE if the absolute Log_2FC was greater than 0.75, Bonferroni corrected p-value < 0.05 and the gene was expressed in at least 10% of the cells that fell within the current knot comparison, in one of the conditions.

10.5.5 GO term analysis of T cell DE genes

Biological process GO term analysis of the lists of DE genes (either unique DE genes or all of them) for the three methods was carried out using the `enrichGO` command of `clusterProfiler`. Background genes were the genes whose expression was more than 5% across the entire integrated T cell dataset. GO terms whose Benjamini-Hochberg corrected p-value was less than 0.01 were identified as significantly enriched GO terms.

10.6 Runtime experiments

Packages were run single threaded on a virtual machine with 70 gigabytes of allotted RAM and 3 GHz clock speed. Runtime experiment code detailing what sections were included for each package when calculating runtime can be found on the TrAGEDy (chapter 1) GitHub, found in the "Code Availability" section .

10.7 Package versions

All analysis and experiments were carried out on R were done with version 4.1.2 except for the Dyngen dataset simulation and the TradeSeq experiments which were carried out on R version 4.1.0. Seurat version 5.0.3, Dyngen version 1.0.5, phateR version 1.0.7, Slingshot version 2.2.1, Single Cell Experiment version 1.16.0, cellAlign version 0.1.0, clusterProfiler version 4.2.2 & TradeSeq version 1.6.0 were used.

For the Genes2Genes experiments, analysis was done on Python version 3.8.16 with the following package versions: genes2genes version 0.2.0, numpy version 1.24.3, pandas version 2.0.3, scanpy version 1.9.3, scipy version 1.10.1, matplotlib version 3.7.1, seaborn version 0.12.2 & scikit-learn version 1.2.2.

11 Chapter 1: Discussion

In this paper we present TrAGEDy, a tool for aligning and comparing single cell trajectories between conditions. TrAGEDy allows the alignment of processes that start and end at different points by identifying the point at which dissimilarity increases from the minimum dissimilarity point. TrAGEDy assesses alignments to remove matches with high transcriptional dissimilarity in portions of the trajectories, allowing the algorithm to additionally capture alignments where the processes diverge and/or converge. TrAGEDy then identifies differences in gene expression over pseudotime between the conditions by taking a sliding window soft clustering approach, allowing it to identify transient changes in gene expression over the shared process.

Across multiple simulated datasets, TrAGEDy can find the underlying alignment between two processes, where cellAlign and G2G fails, due to its ability to prune matches and identify optimal start and end points in the dissimilarity matrix (Fig.9A). Across biological replicates and different conditions of real datasets, TrAGEDy overcomes batch effects to deliver accurate alignments and identify biologically

relevant genes where TradeSeq and Seurat fail.

As of 2019, there were more than 70 TI methods, all of which have varying performance, functionality and constraints (Saelens et al. 2019). Additionally, the chosen method of dimensionality reduction (PCA, UMAP, PHATE, etc) used to generate the embeddings that the trajectory is calculated from also affects the pseudotime results. We chose PHATE for the reduced dimension embeddings as its purpose is to identify transitions in scRNA-seq data. User discretion is advised when choosing the TI and dimensionality reduction method as different methods will give different cell pseudotime values. The method of dimensionality reduction is especially important as many popular methods (like UMAP) have been found to distort cell relationships in the embedding space (Chari and Pachter 2023). Given the variability that can occur before TrAGEDy is applied, a core assumption of TrAGEDy is that the pseudotime must be an accurate reflection of the underlying biological process. Another key consideration when applying TrAGEDy is the density of cells across pseudotime. However, the window TrAGEDy uses to calculate expression at the interpolated points can be weighted to account for regions with few cells. Notably, if cells are present that are separated from the main bulk of the process, like in the WT1 T cell dataset (appendix Fig.9B), this leads to an uneven dissimilarity score between neighboring interpolated points, thus biasing the end alignment. These outlier cells should thus be removed before TrAGEDy is carried out, to ensure a smooth and accurate alignment.

In almost all of the simulated datasets discussed here, TrAGEDy was able to correctly align trajectories, and with greater accuracy than cellAlign and G2G. For cellAlign, the constraints of DTW hamper it from accurately capturing the underlying alignments, while for G2G it may be due to the fact that the final alignment is based off of an aggregate of single gene alignments, rather than directly looking at the global alignment of genes across the cells. Additionally, alignment of real data sets resulted in expected outcomes, including the complete alignment of biological replicates and partial alignments between conditions.

To compare DE results of TrAGEDy we also performed differential gene expression tests using two widely used tools: Seurat and TradeSeq. Although Seurat does not take pseudotime into account when it performs differential expression tests it is possible to compare cells across conditions, something which is not possible in many tools designed to specifically analyse trajectories, including pseudotimeDE and Monocle (Song and J. J. Li 2021 & Trapnell et al. 2014). Low detection of DE genes by Seurat relative to TrAGEDy could be explained by the fact Seurat does not take pseudotime into account when carrying out the differential expression tests.

TrAGEDy identifies many known key processes which are required for *T. brucei* survival in the tsetse fly. While bloodstream forms do not wholly rely on glycolysis for its energy needs, most of its energy is produced through this pathway (Durieux et al. 1991 & Taleva et al. 2023). In contrast to mammals, the tsetse fly is a low glucose environment (Y. Qiu et al. 2018), and the stumpy form is preadapted to suit these low glucose conditions (Grinsven et al. 2009). In the tsetse fly, the stumpy derived PCFs can breakdown amino acids for energy, in particular proline (Evans and R. C. Brown 1972, Weelden et al. 2003, Lamour et al. 2005). The parasite converts proline into glutamate through the action of proline dehydrogenase (*PDH*) and *DP5CDH* with glutamate being further converted, through the action of glutamate-dehydrogenase, into 2-oxoglutarate, which can be used as an energy source in PCFs (Weelden et al. 2003, Mantilla et al. 2017). Only TrAGEDy captures *DP5CDH* as being significantly upregulated in the WT at the end of the shared process, with TrAGEDy and Seurat both capturing glutamate dehydrogenase as being significantly upregulated in the WT towards the end of the shared process. None of the methods captured *PDH* as being DE at any point in the shared process. TrAGEDy thus paints a more complete picture of the proline catabolism pathway, which is active in the preadapted stumpy

forms, than Seurat or TradeSeq.

One of the key unique findings of TrAGEDy was identifying *PAD1* as being significantly upregulated in the WT towards the end of the shared process. *PAD1* is a key marker of stumpy forms, with its absence inhibiting induced differentiation of the parasite from stumpy to the first tsetse fly stage, PCFs (Dean et al. 2009). TrAGEDy also uniquely found that the *AK3* was significantly upregulated in the WT compared to the *ZC3H20* KO. Northern blot analysis has showed that this gene is more highly expressed in PCFs compared to bloodstream forms (Voncken et al. 2013, N.B. the paper refers to AK3 as "AK1") and the kinase has been shown to be important during tsetse fly infection, with parasite infectivity reduced when the gene is knocked out Ooi et al. 2015.

A majority of the genes identified as being DE were upregulated in the WT compared with the *ZC3H20* KO. This difference may reflect the fact that the *ZC3H20* KO has a truncated development, and thus its transcriptome does not change as much as the WTs. What genes we did find to be upregulated in the *ZC3H20* KO were seen with regulatory elements like zinc finger proteins. The two zinc finger proteins (*ZC3H46* and *ZC3H38*) that only TrAGEDy identified as being upregulated in the *ZC3H20* KO have not been identified as being bound by *ZC3H20* (B. Liu, Kamanyi Marucha, and Clayton 2020), but their upregulation in the *ZC3H20* KO cells at the end of the shared process may suggest they are genes whose expression is turned off in the WT due to the effect of *ZC3H20*.

When analysing the T cell development dataset, TrAGEDy captured more significant DE genes than Seurat and TradeSeq. While getting a higher number of DE genes is an important benchmark, it is also important to assess whether these genes fit into the context of the existing literature surrounding the datasets. TrAGEDy uniquely captured many cell-cycle associated genes as significantly upregulated in the WT compared with *Bcl11b* KO. After β -selection, before they begin to rearrange their TCR α chain, T cells undergo a proliferative burst and then become quiescent (Kreslavsky et al. 2012, Hwang et al. 2020). In accordance with this, cell cycle genes are mostly DE and upregulated in the WT in the final two windows (windows 5 and 6). Furthermore, the GO terms that only TrAGEDy captured were mainly associated with cell cycle function. This indicates that TrAGEDy has managed to capture the transient changes in proliferative state as WT T cells pass β -selection and prepare to rearrange their TCR α chain. Furthermore, TrAGEDy uniquely detected an increase in expression of *Cd28* and *Ctla4* in the WT compared to the *Bcl11b* KO. This finding further solidifies the conclusion that the WT T cells are able to pass β -selection while the *Bcl11b* KO are unable to, as *Ctla4* expression increases after T cell activation and an increase in *Cd28* expression has been seen following successful β -selection in maturing T cells (Linsley et al. 1992 & Teague et al. 2010).

TrAGEDy uniquely identified many transcription factors as being upregulated in the *Bcl11b* KO cells. In the original paper, the authors identified a variety of transcription factors whose expression was associated with *Bcl11b* KO fate: for example *Gata3* and *Sox5*, which TrAGEDy also identifies as being upregulated in the *Bcl11b* KO compared with the WT. Another transcription factor only identified by TrAGEDy as significantly upregulated in the *Bcl11b* KO cells at the end of the shared process was *Batf*. Previous research shows that mRNA expression of *Batf* is absent in the DP stages of T cell development (Williams et al. 2001), which occurs after β -selection, suggesting TrAGEDy is capturing the turning off of *Batf* expression as the WT cells transition from DN to DP stages occurs, while the *Bcl11b* KO stays DN. This finding also fits with the fact that the TrAGEDy alignment showed that the *Bcl11b* KO cells did not reach the stages of expressing TCR α chains, implying β -selection was not successful.

In terms of runtime, Seurat performed the best thanks to its optimisation through the Presto algorithm (Korsunsky et al. 2019) but TrAGEDy was not far behind in terms of runtime and performed

better than Seurat in terms of the results returned. The main bottleneck for TradeSeq runtime was fitting the general additive models to each gene. The benchmarking of runtime was done single threaded. TradeSeq can be run in parallel, however the availability of computational power with the capability to run parallel is not widespread. TrAGEDy thus serves as a better alternative requiring less time to run, less resources and provides more biological insightful results than TradeSeq.

Due to their respective algorithms, TrAGEDy, G2G and cellAlign are restricted to aligning linear processes. To overcome this limitation, branching processes could be treated as multiple root-to-leaf linear processes and aligned and analysed in a pairwise fashion, but how accurate this process would be is uncertain as the two paths are not independent of one another. Thus, future work needs to be carried out to identify a method permitting the alignment of complex trajectories in order to better expand the datasets which trajectory alignment can be applied to.

Ultimately, TI represents an approximation of how a developmental trajectory might look and may not represent the actual ordering of the cells through the process. Adding extra temporal information from lineage tracing will help add more biological validity to single cell data (Wagner and A. M. Klein 2020). TrAGEDy could thus take lineage trace-inferred time, rather than pseudotime, as its input to allow the analysis of such datasets, allowing it to remain relevant as more sophisticated single cell sequencing techniques are introduced. Furthermore, the advent of perturb-seq (Dixit et al. 2016) allows researchers to more easily generate and analyse the effect of genetic perturbations on processes (Schraivogel et al. 2020), providing multiple and diverse opportunities for TrAGEDy to be applied.

This chapter demonstrates the ability of TrAGEDy to overcome batch effects between datasets, without the need for data integration, and to reveal the underlying alignment between datasets; identifying shared and dissimilar points across the independent trajectories. With its alignments, TrAGEDy overcomes constraints of current tools which limit the alignment of datasets which may start and end at different points or differ in the middle. TrAGEDy can identify more DE genes than other methods, and in all cases, TrAGEDy can identify more biologically relevant processes enriched in the different conditions. With the advent of perturb-seq methods, the need for TrAGEDy will become more apparent and with lineage tracing it will make the time axis more biological accurate. Thus, this tool increases the capacity to analyse developmental progression using scRNA-seq to reveal meaningful insights into the shared biological processes.

12 Chapter 2: *In vitro* single cell transcriptomic atlas of *T. cruzi* development

The following work is part of a collaborative project with Dr Manu de Rycker, Dr Marta Garcia-Sanchez and Professor Thomas Otto. Dr Marta Garcia-Sanchez carried out all wet lab experiments (with assistance by Dr Juliana Da Silva Pacheco and Dr Luciana De Sousa Paradela), the details of which has been included for information. Text sections and figures in reference to work and generated by collaborates are credited. I carried out the analysis discussed below.

The single cell resolution of scRNA-seq has made it a great tool for capturing transcriptomic data from a breadth of varying cell types and life cycle stages. These collections of expansive scRNA-seq datasets are termed "atlases", and allows researchers to have a fine grained and coarse grained look of a biological system by capturing snapshots of smaller systems embedded in the larger one while allowing global comparisons of these smaller systems. In this chapter we present the first scRNA-seq atlas of the

four main *in vitro* *T. cruzi* life cycle stages. From this data, we define general marker genes for the four life cycle stages and show that the atlas can be utilised to annotate novel *T. cruzi* scRNA-seq datasets. We also provide insights into possible distinct trypomastigote subsets and identify novel marker genes of the metacyclogenesis process.

13 Chapter 2: Introduction

As scRNA-seq has become more widespread and cost effective, increasingly larger and more heterogeneous datasets are being generated, allowing researchers to capture the transcriptome of entire biological systems. These analyses (and the datasets within them) are often termed "atlases", an increasing popular term in modern scRNA-seq analysis. In this context, a scRNA-seq atlas is defined as a collection of curated scRNA-seq datasets which aim to capture the transcriptional heterogeneity of a life cycle, disease, organ, etc. As these atlases capture a broad spectrum of cell types/life cycle stages, they help serve as a reference for future researchers to compare their own data, helping to annotate cells and validate their results.

There are a variety of atlas level studies of mammalian systems, such as Tabula Sapiens (Tabula Sapiens et al. 2022) and Muris (Tabula Muris et al. 2018), which aim to capture the breadth of cells across different tissues in humans and mice, respectively. Atlas level studies also exist for parasites, although the breadth of life cycle stages captures varies. For example, many scRNA-seq datasets exist which look at the parasite in one particular setting i.e. host or vector. For *Schistosoma mansoni*, across multiple papers, scRNA-seq data for the larvae and adult forms of the mammalian stages have been captured, as well as the first intra-vector stage (Diaz Soria, Lee, et al. 2020, Diaz Soria, Attenborough, et al. 2024 & Wendt et al. 2020) however some of the vector stages are still missing. For *T. brucei* many of the life cycle stages have been captured over multiple scRNA-seq datasets with the salivary gland insect stages being captured *in vivo* (Vigneron et al. 2020, Hutchinson et al. 2021 & Howick, L. Peacock, et al. 2022) and the mammalian stages captured *in vitro* (Briggs, Rojas, et al. 2021). No paper has specifically investigated the midgut and proventriculus stages however. Finally, the Malaria Cell Atlas (Howick, Russell, et al. 2019) is the first parasite atlas and the only parasite atlas to capture all the main life cycle stages in one study, making it the most comprehensive atlas analysis of parasites thus far.

There are parasites where these types of analysis do not exist, such as *Leishmania* for which scRNA-seq data only exists for the promastigote stage (Louradour et al. 2022). In the case of *T. cruzi*, microarray data exists for the four main life cycle stages of the parasite (Minning et al. 2009) and bulk RNA-seq data has been generated on epimastigotes, amastigotes and trypomastigotes (Y. Li et al. 2016) and the process of epimastigote to metacyclic trypomastigote transition (metacyclogenesis) (Smircich, Eastman, et al. 2015 & Smircich, Perez-Diaz, et al. 2023). At time of writing, no scRNA-seq data has been published on any of the *T. cruzi* stages.

In this chapter, we present not only the first scRNA-seq datasets of *T. cruzi*, but the first single cell transcriptomic atlas of the complete *in vitro* life cycle of any kinetoplastid. In the atlas, we have captured the four major life cycle stages of *T. cruzi*. Briefly, the *T. cruzi* life cycle is as follows. The host cell infective metacyclic trypomastigotes, are passed into the host through the *T. cruzi* vector, the triatomine. These cells then infect host cells and differentiate into amastigotes. These cells persist inside of the cell for several days before transitioning into trypomastigotes and escape the host cell by causing its lysis. The trypomastigotes are then taken up by the triatomine during a bloodmeal and differentiate into epimastigote forms which then in time differentiate into metacyclic trypomastigotes, restarting the life cycle. We demonstrate that the atlas can be used to accurately annotate query *T. cruzi* scRNA-seq

cells to the different life cycle stages. Finally, utilising the fine resolution of scRNA-seq, we identify trypomastigote subsets at the transcriptomic level, capture the process of metacyclogenesis, and identify many novel genes which may play roles in facilitating the form and function of *T. cruzi* across its life cycle.

14 Chapter 2: Results

14.1 Data generation

The following life cycle stage samples were collected for 10X Chromium V3 3' scRNA-seq by Dr Marta Garcia Sanchez with assistance by Dr Juliana Da Silva Pacheco and Dr Luciana De Sousa Paradela: amastigotes (AMA), either 6, 24 or 120 hours post infection (hpi); epimastigotes (EP); stationary epimastigotes/metacyclic trypomastigotes (SE/MT); and amastigote-derived trypomastigotes (ADT) (Fig.12A). The scRNA-seq datasets samples were made with the following mixes: A mix of amastigotes 120 hours post infection and epimastigotes (EP/AMA₁₂₀), stationary epimastigotes and metacyclics (SE/MT), amastigotes 6 hours and 24 hours post infection (AMA_{6/24}), amastigote derived trypomastigotes (ADT) and a mix (MIX) of SE/MT, EP, ADT and the three different time points of AMA (Fig.12B). Replicates were taken of the MIX and SE/MT scRNA-seq samples, for a total of seven samples.

14.2 Mapping and quality control of *T. cruzi* scRNA-seq data

Cells from the Sylvio X10/1 strain of *T. cruzi* were utilised for all the data generated in this chapter, which is part of DTU I (Herrerros-Cabello et al. 2020). As such the data was initially mapped against the Sylvio X10/1 reference (Franzen et al. 2011) using Cellranger Count. For the MIX 1 dataset, this led to 135 median genes being detected per cell and 2% of reads mapping confidently to the transcriptome (Table 1). The most complete reference genome of a DTU I strain is from the Dm28c 2018 strain (Herrerros-Cabello et al. 2020). To see if the mapping statistics could be improved, the scRNA-seq reads were mapped against the Dm28c 2018 reference (Berna et al. 2018). With this new mapping, the number of median genes increased to 216 and the percentage of reads confidently mapping to the transcriptome increased to 4.7% (Table 1).

Reference Genome	Median genes per cell	No. of cells	% mapping reads to genome	% reads mapping confidently to transcriptome
Dm28c 2018	216	5,464	55	4.7
Sylvio X10/1	135	5,804	15.8	2.0

Table 1: Mapping statistics for the MIX 1 scRNA-seq dataset, when mapped against the Dm28c 2018 and Sylvio X10/1 references

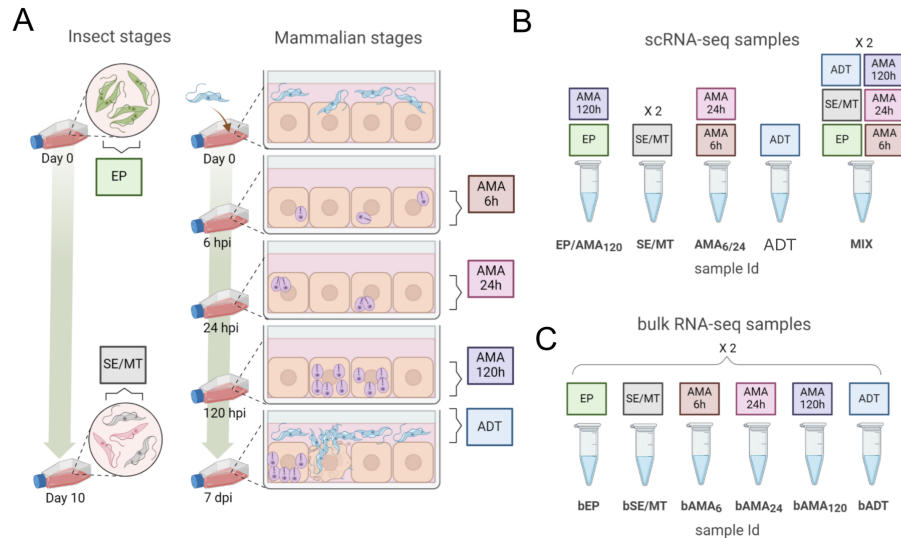


Figure 12: Sampling and library construction

This figure was designed by Dr Marta Garcia Sanchez and the figure legend was edited from a draft written by Dr Marta Garcia Sanchez. Samples collected from *in vitro* cultures: EP: epimastigotes, collected during the exponential growth phase, SE/MT: a mixture of stationary growth phase epimastigotes and metacyclic trypomastigotes, ADT: amastigote derived trypomastigotes and amastigotes isolated from Vero-infected cells at 6, 24, and 120 hours post infection (AMA₆, AMA₂₄ and AMA₁₂₀ respectively). The sampling times for collecting AMA were based on the findings by Li and colleagues (Y. Li et al. 2016). To capture the transition from amastigotes to trypomastigotes, the collecting time of 120 hpi was chosen. The SE/MT sample was collected instead of purifying the metacyclic trypomastigote population, to preserve the natural progression between insect stages, allowing for a more comprehensive analysis of the transcriptome changes that occur during metacyclogenesis. Trypomastigotes were captured 7 days post infection (A). Individual libraries were constructed for SE/MT (two biological replicates) and ADT from single-cell suspensions. Two more libraries were prepared by combining AMA₆ with AMA₂₄ (AMA_{6/24}), and EP with AMA₁₂₀ (EP/AMA₁₂₀) in equal proportion. A second biological replicate was generated by pooling equal numbers of EP, SE/MT, ADT, AMA₆, AMA₂₄ and AMA₁₂₀ in a single-cell suspension (MIX). Two libraries (technical replicates) were sequenced for MIX (B). Two bulk RNA-seq libraries were generated for each of the life cycle stages (C).

Sample	Total reads	Total reads <i>T. cruzi</i>	Median genes (<i>T. cruzi</i>)	% confident mapping to <i>T. cruzi</i> transcriptome	% mapping to <i>T. cruzi</i> genome	% mapping to human genome
AMA _{6/24}	952,950,562	387,850,879	307	1.6	6.3	59.3
EP/AMA ₁₂	682,978,286	678,197,438	928	14.4	43.1	0.7
ADT	729,447,233	729,447,233	762	7.7	30.1	NA
SE/MT 1	271,060,069	271,060,069	298	18.4	44.4	NA
SE/MT 2	251,411,209	251,411,209	421	21.4	57.4	NA
MIX 1	674,582,011	611,171,302	503	10.1	56.2	9.4
MIX 2	739,240,858	663,838,290	422	9.9	56.8	10.2

Table 2: Mapping statistics returned by 10X Cellranger count for the scRNA-seq datasets mapped against the Dm28c 2018 2500bp 3' UTR extended + kDNA + GRCh38 reference

The Dm28c 2018 reference was thus used as the basis for mapping the *T. cruzi* scRNA-seq data. Similar to Briggs and colleagues (Briggs, Rojas, et al. 2021), the Dm28c 2018 transcriptome 3' UTRs were extended by 2500bp and a kDNA reference was appended to the genome. For datasets which did not contain amastigote populations (i.e. SE/MT 1, SE/MT 2 and ADT), the data was mapped against this combined Dm28c 2018, 2500bp 3' UTR extended + kDNA reference (Table 2). For datasets which did contain amastigote populations, the data was mapped to a combined Dm28c 2018, 2500bp 3' UTR extended + kDNA + GRCh38 human genome reference, to account for the fact that the amastigotes are derived from human cells and thus human cell reads will likely be present in the data (Table 2).

The mapping percentages of reads to the *T. cruzi* genome was around ~40% for most of the datasets, with the notable exception being the AMA_{6/24} sample which only had 6.3%. The percentage of reads mapping to the human genome for the AMA_{6/24} sample, however, was 59.3%. This suggests most of the reads in the AMA_{6/24} sample come from the Vero cells use to grow the parasite (Table 2). The low percentage of reads mapping confidently to transcriptome compared to *T. brucei* scRNA-seq runs (Vigneron et al. 2020 & Briggs, Rojas, et al. 2021) may be caused by the high amount of repetitive DNA in *T. cruzi* compared to *T. brucei* due to the presence of many multigene families in the *T. cruzi* genome (Pita et al. 2019). This may also contribute to the lower median genes detected per cell, compared with *T. brucei* runs (Table 2).

To filter out low quality cells, cells were assessed on their total genes captured per cell, percentage of reads that map to the kDNA genes, and (where applicable) the percentage of reads that map to the human genome. After the quality control cutoffs were applied (Table 1), 31,065 parasite cells and 15,321 parasite genes were left in the atlas. The cell expression values were then normalised by the total sum of counts, log+1 transformed, and scaled by a factor of 10,000.

14.3 Integrating the atlas samples

In order to mitigate the influence of batch effects on the seven scRNA-seq samples (Fig.12B), the atlas datasets were integrated together. There are a variety of integration methods available (Luecken, Buttner, et al. 2022). For this chapter Harmony (Korsunsky et al. 2019), Seurat V5 CCA (Y. Hao, T. Stuart, et al. 2024) and BBKNN (Polanski et al. 2020) were considered.

For all integrations, the datasets were first merged together before scaling of the gene expression values and the effect of total UMI count was regressed out. PCA was then carried out on the scaled and regressed expression values, only using expression values of the top 2500 most variably expressed genes. The integration methods were then applied to the PC space, and clustering and UMAP embeddings were generated based on the adjustments made by the integration methods.

As some of the scRNA-seq samples collected (e.g. ADT and AMA_{6/24}) contain only one life cycle stage and the MIX samples contain cells from all the life cycle stages, the presence of these samples across the integrated clustering can be used to assess how well the integration methods are performing. For example, in the Seurat V5 CCA integration the ADT samples cells were present in every cluster (Fig.12C), unlike the Harmony and BBKNN integrations where they were present in two and three clusters respectively (Fig.12A & B). For this reason, Seurat V5 CCA was not utilised for the atlas integration.

The BBKNN and Harmony integrations are similar to one another in terms of the distribution of samples across the clusters (Fig.12A & B). As such both could have been valid choices to integrate the datasets with. BBKNN was ultimately chosen for the final integration to allow myself the opportunity to carry out a scRNA-seq analysis in the Python programming language rather than R (which the other two chapters analyses are coded in).

The first 17 PCs were then used as the basis for BBKNN integration, with the resulting BBKNN graph being used to create UMAP embeddings for the cells and perform Leiden clustering at a resolution of 0.8, generating 10 clusters (Fig.13A).

Our first goal, having created the basis for the atlas, was to assess whether clusters reflecting the different *T. cruzi* life cycle stages were captured in the samples. As many of our samples only contain one or two different cell types, we can use these to assist in the validation of cluster identity, as we would expect the cells from these datasets to only overlap with cells from the MIX datasets. Over 75% of the cells in clusters 0 and 9 came from the ADT sample (Fig.13B, C), indicating that the cells in these clusters are mainly trypomastigotes. Furthermore, around half of the cells in cluster 4 also come from the same ADT sample, with the other half coming from the MIX samples, which also contain trypomastigotes. The cells in clusters 0, 4 and 9 thus represent trypomastigotes. Clusters 1 and 2 both mainly originate from the MIX and EP/AMA₁₂₀ samples (Fig.13B, C), indicating these clusters represent either epimastigotes or late-stage amastigotes. Cluster 5 contains many cells from the AMA_{6/24} (Fig.13B, C) and is spatially adjacent to cluster 1 on the UMAP, but not to cluster 2. Thus, Cluster 5 may represent amastigote cells in the early stages of development from trypomastigotes after host infection, and cluster 2 may represent epimastigote cells.

The remaining clusters contain many cells from the SE/MT scRNA-seq sample. Clusters 3 and 8 are almost completely made up of SE/MT sample cells, meaning they likely represent metacyclic trypomastigotes. Clusters 6 and 7 also contain cells from the MIX and EP/AMA₁₂₀ samples (Fig.13B). Metacyclic trypomastigotes were generated through prolonged culturing of epimastigotes to trigger differentiation, resulting in a mixed population of epimastigotes and metacyclic trypomastigotes. Therefore, clusters 6 and 7 could represent cells that are transitioning from epimastigotes to metacyclic trypomastigotes.

The above analysis, based on our sampling strategy, provides a first indication that the parasites are indeed clustered by life-cycle stage. However, unsuccessful integration of the datasets could lead to incorrect partitioning of cells into different clusters. As mentioned in the introduction, interpreting the distances between points on a reduced dimension projection has been shown to be problematic due to distortions when reducing an ambient space with many thousands of dimensions to just two (Chari

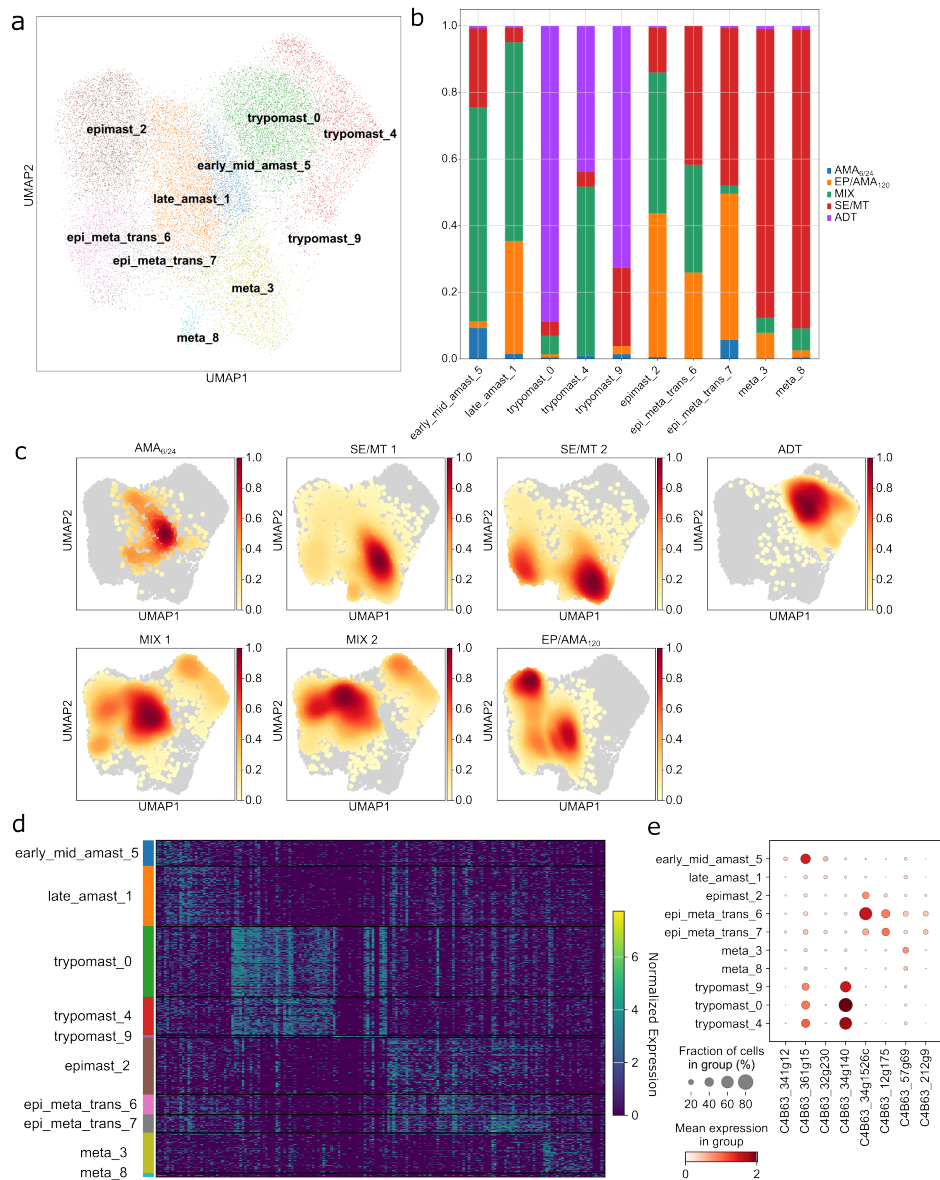


Figure 13: *T. cruzi* atlas overview

UMAP embeddings of the cells of the *T. cruzi* atlas where every cell is coloured by their cluster identity (A). Proportion plot showing the proportion sample origin of the cells across the clusters of the atlas (B). UMAP embeddings of *T. cruzi* atlas cells, split by their sample origin, with each cell coloured by the density of cells in that region of the UMAP. Red means more dense regions of cells, white means less dense (C). Heatmap showing the normalised expression values of the Scanpy derived top 20 marker genes (where available) for each of the clusters in the *T. cruzi* atlas (D). Dotplot showing the expression of selected marker genes for the clusters in the *T. cruzi* atlas. Dots are coloured by the median expression of the gene in each cluster and the size of the dots shows the percentage of cells (in that cluster) that express the gene (E).

and Pachter 2023). To increase our confidence in the biological relevance of the clustering, we sought to associate marker genes for our clusters with known life-cycle markers. The Scanpy derived top 20 marker genes per cluster ($\log_2FC > 0.5$ and Bonferroni corrected p-value < 0.05) (appendix file 14),

where available, were plotted in a heatmap for all cells in the integrated dataset (Fig.13D), as well as violin plots of some selected genes from the marker genes list (Fig.13E).

The following investigation of cluster markers was carried out collaboratively with Dr Marta Garcia Sanchez.

The genes C4B63_34g140, C4B63_147g63 and C4B63_34g324, which show increased expression in clusters 0, 4 and 9 (Fig.13E), encode surface proteins from key multi-gene families associated with trypomastigotes: a trans-sialidase from group IV, a mucin-associated surface protein (MASP), and a member of the Trypomastigote, Alanine, Serine and Valine rich protein subfamily C, respectively. These play different key roles in the biology of trypomastigotes, from the establishment of an effective interaction with the host cell and invasion, to protection against the host immune response (Ralph et al. 2013, Herreros-Cabello et al. 2020). C4B63_32g230 was a marker gene of both clusters 1 and 5, and encodes a member of the amastin glycoprotein family, which is highly expressed in amastigotes (Coughlin et al. 2000). This fits with our assignments of these clusters as amastigotes based on the sampling analysis. Interestingly, some trypomastigote-associated transialidases are also markers for cluster 5 (e.g. C4B63_361g15 and C4B63_51g213), which reflects that this cluster contains cells from the AMA_{6/24} sample, and thus likely contains intermediate stages during differentiation from trypomastigotes to amastigotes. This is substantiated by another marker for this cluster, the lipase encoding gene C4B63_341g12, which is involved in amastigote development (Y. Li et al. 2016). Furthermore, a marker gene of both cluster 4 and 5 is C4B63_32g1409c, which encodes chagasin. Chagasin's primary function is to inhibit and regulate the action of cruzipain, a parasite-specific cysteine peptidase that plays a key role in parasite invasion of host cells (C. C. Santos et al. 2005 & V. C. Santos et al. 2021). As cluster 4 represents trypomastigotes and cluster 5 early amastigotes, the expression of chagasin fits with the biology, as these cells are primed to infect cells or are in the early stages of host cell infection, respectively. These results are thus consistent with the previous conclusion that cluster 5 represents early or mid amastigotes.

Our assignment of cluster 2 as epimastigotes is supported by the marker gene C4B63_34g1526c (Fig.13E), which encodes *TcSMUGL*, a mucin-type glycoconjugate restricted to the surface of *T. cruzi* epimastigotes and is involved in the interaction with the insect vector (Gonzalez et al. 2013). Clusters 3 and 8, which mainly contain cells from the SE/MT sample, show increased expression of C4B63_57g69, encoding a metallo-peptidase, which has been previously shown to be upregulated in metacyclic trypomastigotes in comparison to the other life cycle stages (Minning et al. 2009 & Smircich, Eastman, et al. 2015). These observations support the suggestion that these two clusters represent metacyclic trypomastigotes. Finally, clusters 6 and 7, which we propose represent cells along the epimastigote to metacyclic trypomastigote differentiation axis, are characterised by expression of marker genes C4B63_12g175 and C4B63_212g9 (Fig.13E). These genes are two putative members of the Nodulin-like/Major Facilitator Superfamily, which have been previously associated with the epimastigote to metacyclic trypomastigote transition (Smircich, Eastman, et al. 2015). Taken together, the marker gene analysis supports our cluster allocation, based on our sampling strategy.

14.4 Bulk RNA-sequencing analysis of the *T. cruzi in vitro* life cycle

While the paper by Minning and colleagues (Minning et al. 2009) investigates the four main life cycle stages using microarray, no such paper exists for bulk RNA-seq data. The power of having such an analysis is not only in terms of its novelty; but also using such a resource to validate findings from the scRNA-seq data.

Thus, bulk RNA-seq data for the same life cycle stages and timepoints captured in the scRNA-seq data was generated by the De Rycker lab (Fig.12C) and mapped against the combined Dm28c 2018 2500bp 3' UTR extension + kDNA reference genome (Table 3). The data was analysed with DESeq2 (Love, Huber, and Anders 2014) under default parameters. The positioning of the samples in the PC space was consistent with distinct transcriptomes being present for each of the life cycle stages. The three amastigote samples were localised next to each other. The epimastigotes and metacyclic trypomastigotes were both located in the top left of the PC space and the trypomastigote samples were localised together on the positive end of PC1. Furthermore, the biological replicates of each life cycle stage all localised together (Fig.14A).

Sample	Number of input reads	Uniquely mapped reads %	% of reads mapped to multiple loci	% of reads mapped to too many loci	% of reads unmapped: too short	% of reads unmapped: other	Total alignments (feature-Counts)	% Successfully assigned alignments (feature-Counts)
bAMA ₆ 1	23,475,929	41.88	14.48	9.83	31.79	2.02	21,942,505	38.2
bAMA ₆ 2	20,219,803	46.33	16.26	10.92	24.33	2.17	21,107,762	38.2
bAMA ₂₄ 1	20,498,527	36.14	12.46	5.11	44.57	1.73	16,395,372	38.1
bAMA ₂₄ 2	24,492,333	35.19	12.00	4.52	46.42	1.87	18,848,721	39.5
bAMA ₁₂₀ 1	26,077,005	52.40	17.93	16.22	9.89	3.56	30,285,994	36.4
bAMA ₁₂₀ 2	27,749,110	49.86	17.13	21.31	8.70	2.99	31,157,951	35.0
bADT 1	27,090,531	51.03	20.71	22.63	3.91	1.73	34,960,443	31.4
bADT 2 1	21,882,358	48.82	20.93	22.23	5.18	2.84	27,812,588	31.8
bADT 2 2	22,048,702	43.12	19.93	28.85	5.46	2.64	26,768,688	29.3
bEP 1	25,717,422	54.56	19.39	18.14	4.01	3.90	32,736,362	34.3
bEP 2	25,985,438	56.38	18.93	16.58	3.91	4.20	32,444,213	36.8
bSE/MT 1	22,249,655	59.79	18.83	14.13	4.02	3.23	28,940,501	36.5
bSE/MT 2	28,225,858	64.08	18.66	9.96	4.78	2.52	36,415,986	39.5

Table 3: Mapping statistics returned for the *T. cruzi* bulk RNA-seq datasets mapped against the Dm28c 2018 + kDNA reference

We next identified marker genes for the each life cycle collection point by comparing each point against all other points, extracting genes that had a $\log_2\text{FC}$ greater than 0.5 and a Benjamini-Hochberg corrected p-value < 0.05 . This identifies genes that are markers for each sample, relative to all other samples (Fig.14C, appendix file 16), as well as for each life-cycle stage relative to all other life-cycle stages (i.e. all amastigote samples combined) (Fig.14B, appendix file 15). The scaled, normalized expression values of these markers were plotted on a heatmap, with distinct transcriptomic patterns seen across the life cycle collection points (Fig. 14B), and the three intracellular amastigote timepoints (Fig.14C).

To identify significant marker genes of the life cycle stages, each life cycle stage was compared against the other three stages, with significant genes defined and those whose $\log_2\text{FC}$ was greater than 0.5 and had a Bonferroni corrected p-value < 0.05 . To make the marker genes specific to a given life cycle stage, any genes which overlapped between marker genes lists were removed. This analysis allowed the identification of highly specific marker genes for each life-cycle stage, with 344 genes significantly upregulated in bEP, 510 in bSE/MT, 1381 in bAMA, and 1842 in bADT (appendix file 15). A substantial proportion of the bADT marker genes encode for multigene family surface proteins, including 413 trans-sialidases from groups I-VIII, 320 MASPs, 488 mucins and 44 glycoprotein (GP63s). All of these proteins are known to be highly expressed in trypomastigotes (Roberts et al. 2014). Well-defined markers for the other life cycle stages were also identified, such as amastin (C4B63.32g230) for amastigotes, and mucin *TcSMUGL* (C4B63.34g1450c) for epimastigotes. Interestingly, we were able to identify specific markers for different stages of amastigote development (Fig.14c). Similar to results in the scRNA-seq data, Chagasin (C4B63.32g1409c) were identified as markers of the early amastigote population (bAMA₆), further suggesting that inhibition of cruzipain, or other cysteine proteases, is important at the early stages of amastigote differentiation. Another marker of interest for this population is ATG7 (C4B63.84g61), pointing to a role for autophagy in this population. For the bAMA₂₄ population, marker genes include amastin (C4B63.32g229), three ABC transporters and superoxide dismutase (C4B63.18g258). For the late bAMA₁₂₀, marker genes include one Mucin I (C4B63.63g77) and two Mucin II encoding genes (C4B63.11g418 & C4B63.63g63), a hexose transporter (C4B63.14g3) and two prostaglandin E synthase-2 genes (C4B63.8g300 & C4B63.50g199).

As stated earlier, the bSE/MT sample does not solely contain metacyclic trypomastigotes, but also epimastigotes and transitional forms on the path to differentiation towards metacyclic trypomastigotes. This is reflected in the upregulation of genes involved in the response of *T. cruzi* epimastigotes to pH and oxidative stress, as well as nutritional starvation (Smircich, Perez-Diaz, et al. 2023 & Miranda et al. 2006), such as ascorbate peroxidase (C4B63.135g21), superoxide dismutase (C4B63.9g448), and arginine kinase (C4B63.26g282), along with genes involved in cellular signalling. 3',5'-cyclic adenosine monophosphate (cAMP) serves as a universal second messenger, previously demonstrated as pivotal in mediating metacyclogenesis in *T. cruzi* (Rangel-Aldao et al. 1987). Key components of the cAMP signalling pathway are upregulated in our SE/MT sample, including cAMP phosphodiesterase A (C4B63.4g243) alongside two protein kinase A catalytic subunits (C4B63.2g156 & C4B63.4g242).

When treating the AMA samples as one, we identified several surface proteins, including five amastins, 15 GP63 surface proteases, a MASP (C4B63.8g96), and eight trans-sialidases, which have all been implicated with the establishment of intracellular infection in *T. cruzi* (Herrerros-Cabello et al. 2020, Alvarez et al. 2008 & Miranda et al. 2006). Additionally, we identified genes involved in metabolism, predominantly involving lipid metabolism, alongside transporters engaged in scavenging nucleosides and other essential nutrients. Among these, we identified four nucleoside transporters, one polyamine transporter (C4B63.2g228), two folate/pteridine transporters (C4B63.7g126 & C4B63.21g273), and nine amino acid transporters. This is in line with the need for intracellular amastigotes to obtain key metabolites from the host cell (Caradonna et al. 2013). *T. cruzi* exhibits auxotrophy for polyamines, folates, and pteridines

(De Pablos et al. 2011 & Kulkarni et al. 2009) and import of amino acids is also critical for survival (Kimuda et al. 2019).

Among the markers for epimastigotes, we also identified genes involved in cell division, energy production in the mitochondria, and metabolism, with a predominance of genes involved in amino acid metabolism (appendix file 15). Additionally, several genes were identified which have been reported in previous studies as epimastigote markers such as arginase (C4B63_4g462) and proline racemase (C4B63_52g138) (Y. Li et al. 2016 & Smircich, Eastman, et al. 2015).

As we have captured both bulk and single cell RNA-seq data from the *in vitro* life cycle, we assessed whether the expression of marker genes derived from our scRNA-seq clusters match up with their respective bulk RNA-seq clusters (Fig.14D). The markers of the scRNA-seq clusters were generally highest expressed in their equivalent bulk RNA-seq samples, across both biological replicates. Some marker genes were highly expressed across multiple life cycle bulk RNA-seq samples, with the markers of single-cell clusters 6 and 7 being highly expressed in both the bEP and bSE/MT bulk RNA-seq samples (Fig.14D). This supports the suggestion that clusters 6 and 7 represent epimastigote to metacyclic trypomastigote transitional cells. Markers for cluster 9, which consists of a relatively small number of cells (Fig.13A), mainly originating from the ADT sample (Fig.13B) and showed less clear alignment with the bulk data, with some genes expressed highly in amastigotes, trypomastigotes and epimastigotes (Fig.14D).

From the analysis of the bulk RNA-seq and scRNA-seq datasets, key markers of the four major life cycle stages are identified, including key metabolism genes for amastigote survival in host cells, proteins required for trypomastigote invasion of host cells, stress responses in epimastigotes and intracellular signalling pathways in metacyclic trypomastigotes.

14.5 The *T. cruzi in vitro* atlas as a tool for annotating datasets

The power of a scRNA-seq atlas is creating a resource that future researchers can utilise to compare with or bolster their own data when investigating their research questions. As such, one of our goals was to create a resource that the wider community can utilise to give confidence in their findings. One of the primary uses of a scRNA-seq atlas is as a reference when predicting cell type labels on an unlabelled dataset. First, we tested whether the atlas could be used to correctly annotate a subset of cells from the atlas itself. Using scPoli (De Donno et al. 2023), a model was trained on subsets of the atlas and used to predict the life cycle stage labels of the rest of the dataset. The model is trained by taking the gene expression values of the training data cells and identifying some weights which embeds the data into a latent space (akin to dimensionality reduction) and some other weights which can take the latent space data and reconstruct it back into the original gene expression data. The weight values are optimised by minimising the difference in expression values between the original expression data and the reconstructed expression data. Predictions are then made by embedding the testing data cells, using the same weights generated from the training, and picking the cell label from the closest training cell in the latent space.

	0.05 (1)	0.05 (2)	0.25 (1)	0.25 (2)	0.5 (1)	0.5 (2)	0.75 (1)	0.75 (2)	0.9 (1)	0.9 (2)
early_mid_amast_5	0.7985	0.7396	0.7336	0.7286	0.7131	0.7255	0.6309	0.6157	0.4259	0.4700
epi_meta_trans_6	0.7528	0.7660	0.7607	0.7505	0.7497	0.7595	0.7039	0.6492	0.4660	0.4439
epi_meta_trans_7	0.5854	0.6144	0.5899	0.6984	0.6004	0.6261	0.5367	0.5685	0.3923	0.3135
epimast_2	0.8738	0.8549	0.8561	0.8470	0.8317	0.8433	0.8241	0.8118	0.7734	0.7867
late_amast_1	0.7968	0.7724	0.7561	0.7868	0.7284	0.7408	0.6919	0.6625	0.5734	0.5551
meta_3	0.8520	0.8247	0.7894	0.7845	0.7878	0.7768	0.7552	0.7353	0.7303	0.7452
meta_8	0.7879	0.7500	0.7872	0.7914	0.7447	0.6624	0.6250	0.6577	0.5962	0.6062
trypomast_0	0.7574	0.7776	0.7513	0.7547	0.7451	0.7494	0.7321	0.7325	0.7197	0.7095
trypomast_4	0.7204	0.7453	0.7163	0.7295	0.6969	0.7005	0.6839	0.6783	0.6695	0.6455
trypomast_9	0.6364	0.6957	0.6818	0.6863	0.7333	0.4136	0.6333	0.6047	0.4076	0.2938

Table 4: **Table of f1-scores for each cluster and proportion of the atlas used as testing data**

Each column represents a different proportion (to the left of the underscore) of the atlas used as testing data for scPoli and what replicate (within the brackets) it was.

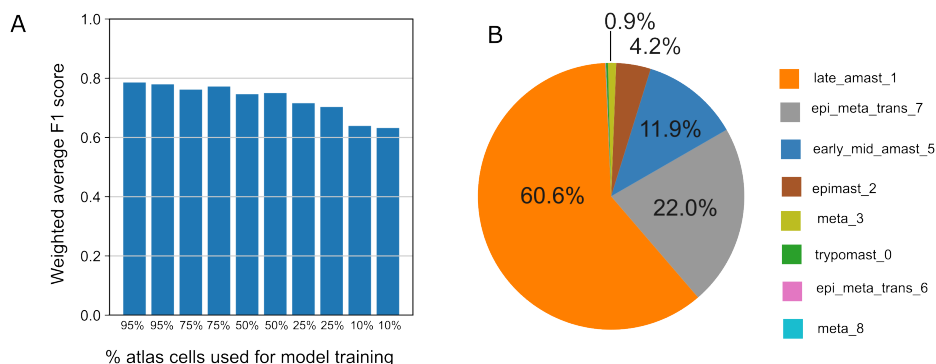


Figure 15: **Annotation results when training scPoli on the *T. cruzi* atlas dataset**

Barplot showing the weighted average F1 score of scPoli (trained on different percentage subsets of the atlas) predictions of cell type annotations on the testing data (A). Pie chart showing the scPoli cluster ID predictions for the 48 hours post infection derived amastigotes when trained on the entire *T. cruzi* atlas (B).

As model training can be computationally expensive, we trained the scPoli model using increasingly smaller training data sizes to find an optimal trade-off between annotation accuracy and training size. To test how well the testing data was annotated, a weighted average f1-score was used. When the f1-score is close to 1, it represents that the model has perfectly annotated the cells in the datasets, without any false positives or false negative cell label assignments. The closer to 0 the f1-score, the opposite is true. Across all of the subsets tested, the weighted average f1-score did not drop below 0.6 (Fig.15A). Furthermore, the 95% training data subset only returned a slightly higher weighted average f1-score than the 50% training data subset, showing the atlas can accurately label the cells with only half the atlas (Fig.15A). Variation was observed in the classifying accuracy of the different clusters. The epimast_2 and meta_3 clusters had f1-scores of ~ 0.84 when using 95% of the data to train the model on and ~ 0.76 when only 5% was used (Table 4). Conversely, the epi_meta_trans_7 cluster f1-score with the 95% training subset was ~ 0.58 and decrease to ~ 0.35 for the 5% training subset (Table 4).

As the testing and training datasets are derived from cells which come from similar sequencing runs and were generated in similar ways, it represents a best-case scenario for predicting life cycle stage. As future researchers' data may be generated through different means or under different conditions, we next aimed to see whether the atlas could be used to correctly annotate cells that were sampled independently from those in the atlas itself. We generated a separate scRNA-seq dataset of amastigotes harvested 48 hpi and used an scPoli model trained on the entire *T. cruzi* atlas to predict life cycle stage labels. As these amastigotes lie in between the mid and late stages post-infection amastigotes, we considered a cluster annotation of late_amast_1 or early_mid_amast_5 as correct and anything else as incorrect. Using this approach, 72.5% of cells were correctly annotated as amastigote (Fig.15B), suggesting that the atlas is suitable for supervised labelling of novel *T. cruzi* scRNA-seq data.

14.6 Cell cycle analysis of the *T. cruzi* atlas

The replicative status of the different life cycle stages has been well established in the literature, with metacyclic trypomastigotes and trypomastigotes being non-replicative, and amastigotes and epimastig-

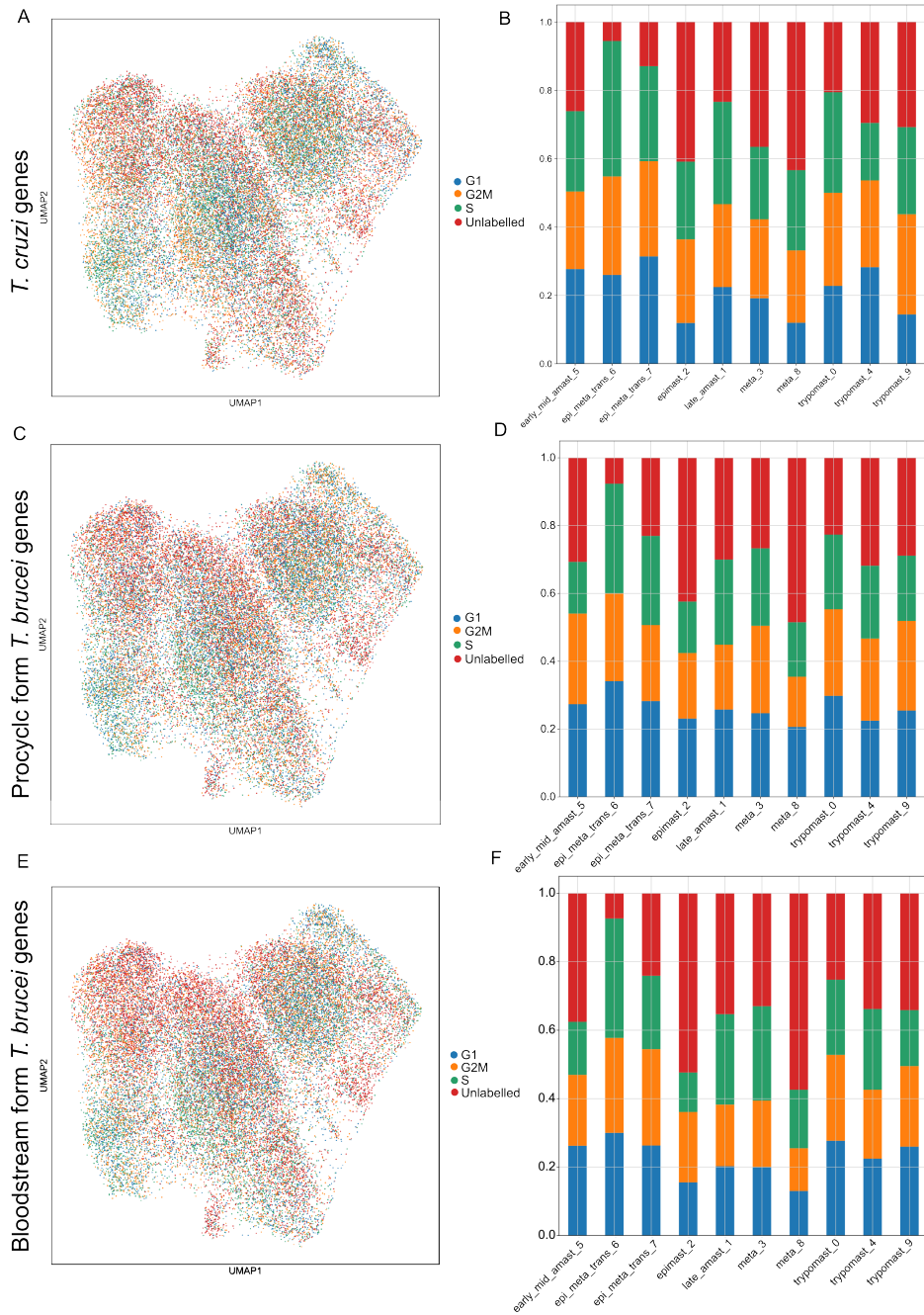


Figure 16: **Cell cycle phase annotation of the atlas cells**

UMAP embeddings of the atlas cells coloured by their cell cycle phase using the Chávez and colleagues (Chávez, Urbaniak, et al. 2021) *T. cruzi* (A), PCF *T. brucei* (C) or BSF *T. brucei* cell cycle genes (E).

Associated proportion plots showing the proportional makeup of each atlas cluster, in terms of cell cycle phase, for the *T. cruzi* (B), PCF *T. brucei* (D) and BSF *T. brucei* (F) cell cycle genes

otes being replicative (Taylor et al. 2020). We therefore tested whether information on cell cycle phase could be extracted from the scRNA-seq data. We took the CL Brener Esmeraldo-like genes that Chávez and colleagues (Chávez, Eastman, et al. 2017) identified as peaking in the different cell cycle phases of epimastigotes, found their syntenic orthologs in the Dm28c 2018 reference using TriTrypDB (Shanmugasundram et al. 2023), and used these genes to assign cells with a cell cycle phase label using a method

outlined by Briggs and colleagues (Briggs, Marques, et al. 2023). For all the clusters, very few of the cells were annotated as "unlabelled", indicating that they were present in a detectable cell cycle stage. However, in terms of the clusters that should contain non-replicating cells, ~80% of the cells in the trypomastigote clusters were labelled in terms of cell cycle phase and ~60% of the cells in the metacyclic trypomastigote clusters are also labelled (Fig.16A, B). For the replicating life cycle phases, a third of the cells in the epimastigote cluster were 'unlabelled' while the amastigote clusters had ~20% of cells being 'unlabelled' (Fig.16A, B).

As many of the clusters that should not contain replicating cells had predicted replicating cells, and some of the clusters that should be replicating have many cells which are not labelled as such, we wanted to rule out that the lack of gene expression resolution inherent in scRNA-seq might lead to this disparity. We did this by checking the expression of the cell cycle marker genes (40 random marker per cell cycle phase) in our bulk RNA-seq to see if they matched up with the underlying biology. Across the different cell cycle phase genes and bulk RNA-seq samples, the expression levels of the genes varied, with no sample having the highest expression values for all the cell cycle genes (Fig.13A, B & C). The cells were then annotated by cell phase using PCF and BSF *T. brucei* cell cycle marker genes, identified by Briggs and colleagues (Briggs, Marques, et al. 2023). These results were almost identical to the ones generated using the Chávez derived cell cycle genes (Fig.16C, D, E & F). These results highlight an inability to utilise scRNA-seq data to accurately annotated the cell cycle phase of *T. cruzi* cells, despite this being possible with *T. brucei* and *Leishmania*.

14.7 Characterising trypomastigote heterogeneity

We identified three clusters of trypomastigotes in the atlas, with two of them being the largest and smallest clusters (in terms of number of cells) of the atlas. The smallest, trypomast_9, only amounts to 208 cells and thus interpreting the marker genes of these cells may be prone to outlier effects. The trypomast_0 and trypomast_4 clusters, however, are larger and thus comparisons can be made without the worry of outlier effects. To identify transcriptomic differences defining these two trypomastigote clusters, we identified all the significantly DE genes ($\text{Log}_2\text{FC} > 0.5$, Bonferroni corrected p-value < 0.05) between the two clusters. 73 genes were identified as being significant markers of trypomast_4, while trypomast_0 had 433 genes (appendix file 17). Marker genes of cluster 0 included a variety of metabolism-associated genes such as L-threonine metabolism enzymes L-threonine 3-dehydrogenase (*LDH*) (C4B63.42g129) and 2-amino-3-ketobutyrate coenzyme A ligase (*KBL*) (C4B63.32g217, C4B63.32g214) (Schmidt et al. 2001), and cytochrome oxidase subunits (C4B63.43g134, C4B63.41g212 & C4B63.12g334). Trypomast_4, on the other hand, was defined by such markers as GP63 protease and trans-sialidases. While some of the marker genes of trypomast_0 were also trans-sialidases, only 7 trans-sialidases were markers of trypomast_0, compared to 32 in trypomast_4. Two mRNA binding proteins were also found to be markers of the trypomast_4 cluster, specifically C4B63.13g223 and C4B63.16g295, whose syntenic orthologs in *T. brucei* are *ZC3H32* and *RBP10* respectively. These results seem to suggest that the trypomastigote subsets exist with differing cell states: one with high metabolism associated genes and one with host cell invasion associated genes, with the former possibly representing a reservoir of trypomastigotes for future infections and the latter representing cells which are primed to infect host cells

14.8 Identifying markers of *T. cruzi* metacyclogenesis

scRNA-seq provides a key advantage over RNA-seq when studying cells undergoing transitions between different life cycle stages. In bulk RNA-seq, cells from different points of the development process can be present in the same sample and in the input their transcriptomes will be additive, making it difficult to see what transcriptomic signal is from which cell. We can see this lack of resolution when looking at the expression of our putative epimastigote to metacyclic trypomastigote transitional clusters markers on our bulk RNA-seq data (Fig.14D). Due to the single cell nature of scRNA-seq, this issue does not occur. We sought to identify genes that may play a key role in the transition processes captured in the data. Of the *T. cruzi* life cycle transitions, only amastigote to trypomastigote and epimastigote to metacyclic trypomastigote were carried out for the cells which went into the scRNA-seq and bulk RNA-seq datasets. This, however, does not mean that the transition processes were captured in the data. Clusters which contain both SE/MT and EP/AMA₁₂₀ sample cells were contained within the data, hinting that the epimastigote to metacyclic trypomastigote transition could have been captured. No clusters containing a mix of AMA or ADT sample cells were found in the integrated dataset. Furthermore, the main transcriptomic variation in the bulk RNA-seq datasets was between the ADT samples and every other life cycle stage. These results suggest that the amastigote to trypomastigote transition has not been captured and that the reason it wasn't captured was due to the unique transcriptome of trypomastigotes, compared with the other life cycle stages.

To capture the transition between the epimastigote and metacyclic trypomastigote stages, we utilised the trajectory inference (TI) package Partition-based Graph Abstraction (PAGA) (Wolf, Hamer, et al. 2019) on a subclustered version of the atlas, containing cells from the "epimast_2", "meta.3", "epi_meta_trans_6" and "epi_meta_trans_7" clusters (Fig.17A, B, C). PAGA functions by assigning pseudotime values to cells based on their connections to other cells in a nearest neighbour graph (built from the UMAP embeddings). TradeSeq (Van den Berge et al. 2020) was used to capture genes that were

DE across the epimastigote to metacyclic trypomastigote transition, with 923 genes identified as being DE across the differentiation process (appendix file 18). To see at what point in the differentiation these genes were associated with, we took the smoothed expression of the 923 DE genes and ordered them by their peak expression in pseudotime (Fig.17D). From this, we can predict genes that peak in expression at the early, middle and late points of differentiation. These three points were defined from the smoothed expression pseudotime axis, with genes that were highly expressed in the early or early and middle of the axis being defined as early, the genes highly expressed in the middle of the axis defined as middle and the genes that were highly expressed late or late and middle of the axis defined as late. 231 genes displayed expression that peaked early in the transition, 571 genes peaked during the middle stages and 121 genes peaked at the end of the differentiation process (appendix file 18), providing a rich dataset to study metacyclogenesis.

The expression of known marker genes for epimastigotes and metacyclic trypomastigotes changed as expected. The expression of L-threonine catabolism enzymes LDH (C4B63_42g129) and KBL (C4B63_32g217) (Schmidt et al. 2001) significantly changed over metacyclogenesis and peaked early in the process (Fig.17E). In heme-cultured epimastigotes, the parasites showed high consumption of L-threonine (Silva Dias Vieira et al. 2023), suggesting L-threonine metabolism may be important to epimastigotes.

Proline racemase (C4B63_52g138) was also found to be highest expressed in the early stages of metacyclogenesis (Fig.17E). This protein exists as two isoforms (Proline racemase A and B), with both being important in metacyclogenesis. Their overexpression leads to increased numbers of mammalian-infective metacyclics (Chamond et al. 2005). The two isoform proteins are expressed in different stages - A in stationary epimastigotes, and B in metacyclic trypomastigotes (Reina-San-Martín et al. 2000) - and have high sequence similarity, making them hard to disentangle at the transcriptomic level. Nevertheless, their association with driving metacyclogenesis is consistent with their upregulation at the beginning of metacyclogenesis.

Next, we focused our analysis on genes whose expression peaked during the middle of the transition. The expression of C4B63_34g1526c, which encodes a mucin *TcSMUG-L* surface proteins, peaked during the middle phase (Fig.17E). In line with this profile, *TcSMUG-L* is known to be restricted to the surface of epimastigotes, where it is thought to play a role in attachment of epimastigotes to the insect midgut epithelium (Gonzalez et al. 2013 & Urban et al. 2011). Several DNA repair proteins, namely *RAD51*, *RAD50* and *RAD54* (C4B63_6g507, C4B63_6g508 & C4B63_28g105) were also found to peak in expression during the middle phase of metacyclogenesis (Fig.17E).

Finally, we focused on genes whose expression peaked at the late phase of the transition. Known markers of metacyclic trypomastigotes, the surface protease GP63 (C4B63_16g183) and the cruzipain inhibitor Chagasin (C4B63_32g1409c), peaked late in the metacyclogenesis process (Fig.17E). The up-regulation of these enzymes late in the transition may reflect the parasite turning on expression of genes which regulate infection; preparing for their transfer into the host. Finally, the enzyme citrate synthase (C4B63_4g184) was also associated with the end of metacyclogenesis (Fig.17E), with citrate synthase activity shown to be significantly higher in metacyclic trypomastigotes compared to epimastigotes (Adroher, Osuna, and Lupiañez 1988).

To assess the validity of the markers as being important in the epimastigote to metacyclic trypomastigote transition, we looked at the expression of the markers in the bulk RNA-seq data. Interestingly, while the genes that defined the early and late phases of differentiation were highest expressed in the epimastigote and metacyclic trypomastigote bulk RNA-seq datasets, respectively, no clear pattern was seen for the genes identified as peaking in the middle of the process (Fig. 17F). This reveals an inherent

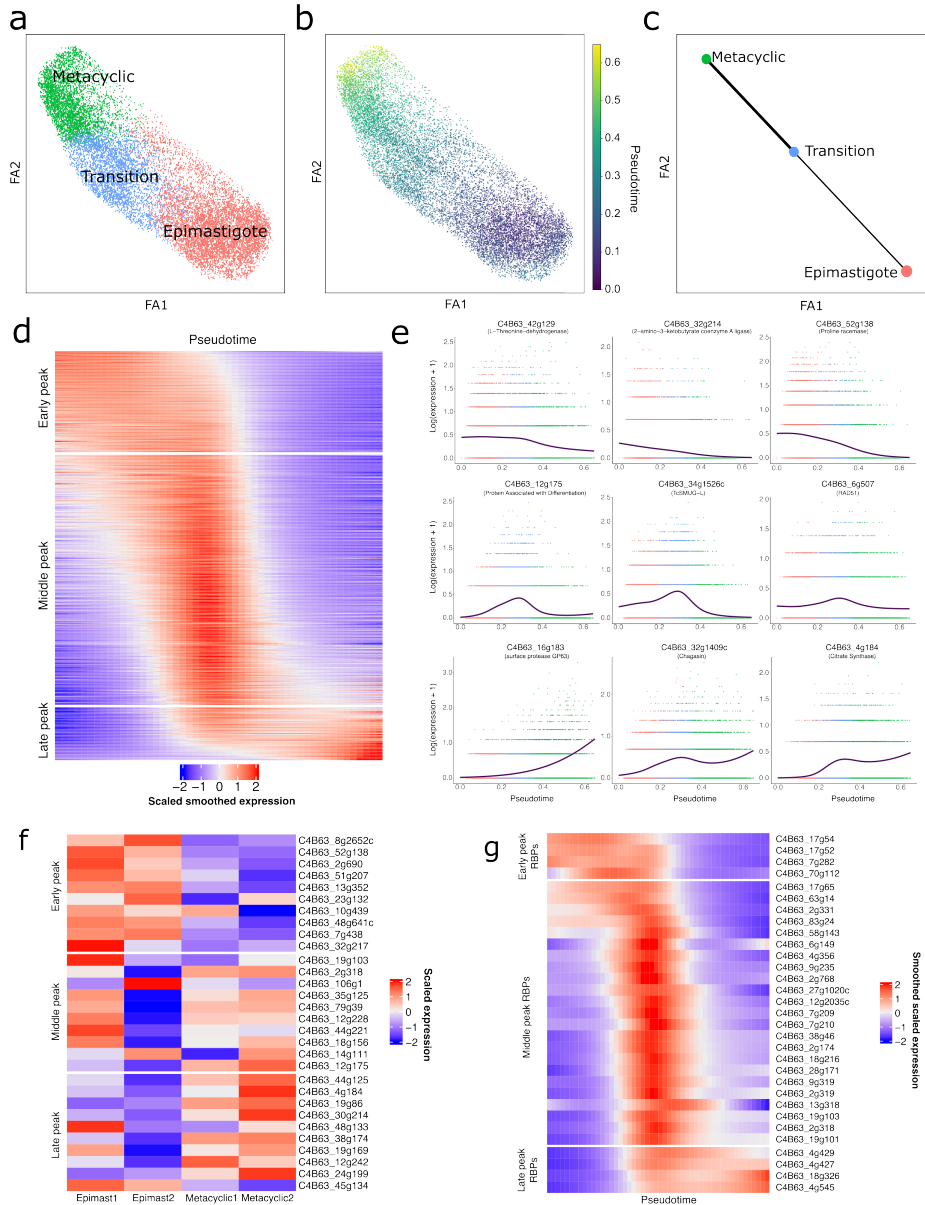


Figure 17: Trajectory analysis of epimastigote to metacyclic trypomastigote transition

FA2 embeddings of the epimastigote, metacyclic trypomastigote, and transitional cells coloured by cluster identity (A) and PAGA derived pseudotime (B). PAGA graph of the FA embeddings, with the dots representing the epimastigote, metacyclic trypomastigote and transitional clusters. The thickness of the lines shows how connected the clusters are, as determined by PAGA (C). Heatmap showing the scaled smoothed expression values (of 50 sampled points across pseudotime) for the genes whose expression significantly changed over the metacyclogenesis axis. Genes are ordered along the Y axis by their peak expression across the metacyclogenesis process (D). Plots showing the log_{1p} transformed expression of selected genes whose expression significantly changed over metacyclogenesis pseudotime. Cells are coloured by their cluster identity (as in A) and the lines represent the smoothed expression of genes across pseudotime (E). Heatmap showing the scaled, normalized expression of the top 10 genes (in terms of Log₂FC) for the early, mid and late peaking genes, whose expression significantly changed over metacyclogenesis pseudotime, on the epimastigote and metacyclic trypomastigote bulk RNA-seq samples (F). Heatmap showing the scaled smoothed expression values (for 50 sampled points across pseudotime) for the RNA binding protein encoding genes whose expression significantly changed over the metacyclogenesis axis. Genes are ordered along the Y axis by their peak expression across the metacyclogenesis process (G)

advantage scRNA-seq has over bulk RNA-seq, in that the granularity of results is much greater. Thus, scRNA-seq is able to identify these transitional clusters and extract their gene signatures while bulk RNA-seq fails, as it is a mixture.

Unlike the bulk RNA-seq samples for the amastigotes, multiple time point samples were not taken as epimastigotes transitioned into metacyclic trypomastigotes. We thus took the bulk RNA-seq data from Smircich and colleagues (Smircich, Perez-Diaz, et al. 2023), who looked at the epimastigote to metacyclic trypomastigotes transition across four time points: Day 7, Day 14, Day 21 and Day 28 of culturing the parasites without nutrient replenishment. As their analysis was carried out on data mapped against the CL Brener Esmeraldo-like v5.0 reference, we remapped their data against our combined Dm28c 2018 2500bp 3' UTR extended + kDNA reference genome and analysed the data with DESeq2, as we did with our bulk RNA-seq data. Remapping of the data did not affect the spatial arrangement of samples within the PCA embeddings compared with the original mapped data (appendix Fig.14A). We visualised the expression of our metacyclogenesis-associated genes in the remapped Smircich data. The majority of the marker genes across the three clusters were highest expressed in the day 7 sample, when the cells should still be epimastigotes in exponential growth phase, although some scRNA-seq markers of the metacyclic life cycle stage were higher expressed at final two time points (appendix Fig.14B). This further highlights the ability of scRNA-seq to capture and clearly visualise gene expression changes that happen over differentiation, where bulk RNA-seq fails due to its lack of granularity.

Given the polycistronic gene expression nature of kinetoplastids, much of the regulation of gene expression is done through affecting mRNA stability and half-life. RBPs are trans-acting factors that engage with cis-acting regulatory sequences in the untranslated regions of mRNAs, mediating mRNA processing, stabilisation, subcellular localisation, and translation efficiency (Romagnoli et al. 2020). Increasing evidence suggests that the expression levels of specific RBPs are closely associated with the regulation of subsets of mRNAs, influencing parasite differentiation in both *T. cruzi* and *T. brucei* (Y. Li et al. 2016, Sabalette et al. 2023). We therefore examined our pseudotime dataset DE list for other RBPs with known associations with differentiation. C4B63.19g103 peaked during the middle phase of the epimastigote to metacyclic trypomastigote transition (Fig.17G). The CL Brener Esmeraldo-like ortholog of C4B63.19g103 (TcCLB.507093.220, *TcUBP1*) has been shown to play a role in mediating the differentiation of *T. cruzi* epimastigotes into metacyclic trypomastigotes under nutritional starvation conditions (Sabalette et al. 2023, Romaniuk, Cervini, and Cassola 2016). Overexpression of *TcUBP1* leads to upregulation of genes encoding surface proteins and alterations in their subcellular localisation within replicative parasites, resulting in increased infectivity of trypomastigotes Sabalette et al. 2023. As well as *TcUBP1*, three other putative RBPs (C4B63.2g318, C4B63.83g24 and C4B63.18g216) also peaked in expression during the middle of metacyclogenesis. These three genes were previously shown to be upregulated in epimastigotes, when compared to amastigote-derived trypomastigotes, and amastigotes (Tavares et al. 2021).

C4B63.4g545 showed increased and sustained expression over pseudotime; peaking in the late phase of the transition (Fig.17G). The gene encodes *RBP4*, whose ortholog in CL Brener non-Esmeraldo-like (TcCLB.508901.20) was expressed higher in metacyclic trypomastigotes in a comparative microarray transcriptome analysis of the four major life cycle stages of *T. cruzi* (Minning et al. 2009). The expression of C4B63.18g326, encoding the zinc-finger protein *ZFP1*, also increased late in the transition process. A role for *ZFP1* in *T. brucei* BSF differentiation has been suggested (Hendriks and Matthews 2005), and upregulation of this RBP in the stationary stage of *T. cruzi* epimastigote growth curve has been previously described (C. Santos et al. 2018).

All the RBPs presented thus far have known roles in metacyclogenesis. Next, we searched the list of

genes significantly associated with metacyclogenesis to identify RBPs that have not been described as associated with metacyclogenesis in *T. cruzi*. C4B63.27g1020c encodes a predicted RBP which peaks in the middle of metacyclogenesis in our data (Fig.17G) and whose CL Brener Esmeraldo-like syntenic ortholog has been associated with late-stage amastigotes (Belew et al. 2017). Interestingly, however, the syntenic ortholog of this gene in the *T. brucei* TREU927 reference (Berriman et al. 2005) encodes *RBP5*, a gene whose expression is upregulated in the mutated *RBP6* overexpression model of *T. brucei* (which will be discussed in more detail in chapter three) insect development (H. Shi, K. Butler, and Tschudi 2018). At the start of metacyclogenesis, we identify a peak in expression of an RNA helicase (C4B63_7g282) whose ortholog in *T. brucei* (Tb927.4.1500) is part of the mitochondrial RNA binding complex 1, which edits mRNAs. When expression of Tb927.4.1500 is repressed with RNAi, PCF *T. brucei* are no longer viable, hinting at an essential role for the gene in the insect stages (Hashimi et al. 2009).

Metacyclogenesis can be induced in epimastigotes through starvation, as done in this paper. During the middle phase of metacyclogenesis, we identified C4B63_58g143 peaking in expression (Fig.17G). The syntenic ortholog of this gene in the TREU927 *T. brucei* reference (Tb927.10.2370) encodes a protein which is significantly enriched in starvation granules in *T. brucei* PCFs (Kafkova et al. 2018). The upregulation of this gene in the transitional middle phases of metacyclogenesis, where the parasite is starved, may suggest this gene is similarly associated with starvation in *T. cruzi*. Our data thus supports the roles of RBPs in facilitating life cycle transitions in *T. cruzi*.

15 Chapter 2: Materials and Methods

15.1 Sample collection

The samples were generated by Dr Marta Garcia-Sanchez with assistance by Dr Juliana Da Silva Pacheco and Dr Luciana De Sousa Paradela of the University of Dundee. The "Sample Collection" subsection within this section (written by Dr Marta Garcia-Sanchez and edited by Dr Manu de Rycker of University of Dundee) details how the biological samples were generated.

15.1.1 Parasite culture

T. cruzi parasites from the Silvio strain (MHOM/BR/78/Silvio; clone X10/1/7) subclone A1 were used in this study (Roberts et al. 2014).

Epimastigotes were grown at 28°C in RTH/FBS [RPMI 1640 medium (Sigma, St Louis, MO, USA) supplemented with 0.4% (w/v) trypticase peptone (BD, Le Pont de Cleix, France), 25 µM hemin (Sigma, St Louis, MO, USA), 17 mM Hepes (Formedium LTD, Hunstaston, UK) pH 7.4, and 10% (v/v) heat-inactivated foetal bovine serum (FBS, Gibco, Paisley, UK)]. The cultures were maintained in exponential growth by continuous passages of 3×10^5 cells/ml every three days.

Metacyclogenesis was induced by nutritional starvation as previously described (MacLean et al. 2018). Briefly, epimastigotes were transferred to a culture flask containing fresh RTH/FBS to a density of 3×10^5 cells/ml and were maintained at 28°C without media change. The relative percentage of metacyclic trypomastigotes to stationary phase epimastigotes was determined by microscopic examination of parasites. On day 10 of culture, the highest percentage of parasites that displayed a metacyclic trypomastigote-like shape (~50%) was observed, aligning with findings reported for the Medellin strain (TcI) (Garcia-Huertas

et al. 2023).

Intracellular amastigotes were grown in Vero cells, as described by MacLean and colleagues (MacLean et al. 2018). In brief, metacyclic trypomastigote-rich cultures were incubated with a monolayer of Vero cells (ECCAC 84113001) at a multiplicity of infection (MOI) of 10 overnight at 37°C 5% CO₂ in Minimum Essential Medium (MEM, Glutamax™ Supplement) (Gibco, Paisley, UK) supplemented with 10% FBS. After overnight infection, extracellular parasites were removed and the Vero cell monolayer was washed three times with serum-free MEM, followed by addition of fresh MEM 10% FBS. Infected flasks were maintained at 37°C 5% CO₂ and the medium was replaced at 72 hours post-infection (hpi). At seven days post-infection (dpi) tissue-culture derived trypomastigotes emerging from infected Vero cells were harvested and used to set-up new infections at MOI 1.5.

15.1.2 Sample collection

Epimastigotes (EP) were collected in the exponential growth phase, from 2 days-old cultures initiated with 3×10^5 cells/ml. A mixture of stationary phase epimastigotes and metacyclic trypomastigotes (SE/MT) was obtained from a 10 days-old culture initiated with 3×10^5 epimastigotes/ml. Amastigote derived trypomastigotes (ADT) were harvested from the supernatant of infected Vero cells at 7 dpi. EP, SE/MT and ADT were put in falcon tubes and left undisturbed for 1 h (at 28 °C for EP and SE/MT, and at 37 °C 5% CO₂ for ADT). To avoid collecting dead parasites and co-purified Vero cells, the supernatant, containing viable and motile cells, was collected. Parasites were pelleted and washed once in ice-cold phosphate-buffered saline (PBS) by centrifugation at 1,350 x g for 5 min at 4 °C.

Amastigotes (AMA) were obtained from Vero cells infected with ADT and washed at 6 hpi as indicated above. At the time points selected (6, 24, and 120 hpi), intracellular AMA were isolated from the host cells following the protocol described by Dumoulin and colleagues (Dumoulin et al. 2020), with minimum modifications. MOI 1.5 was used for 120 hpi samples and MOI 10 for 6 and 24 hpi samples. Infected monolayers were washed with ice-cold PBS and scraped into 3 ml PBS. Infected cells suspensions were lysed using the gentleMACSTM dissociator with gentleMACSTM M tubes (Miltenyi Biotec, Bergisch Gladbach, Germany) and the “Protein” protocol. The ensuing lysates were passed through a PD-10 desalting column (Cytiva, Little Chalfont, UK) equilibrated with PBS to remove debris. AMA were eluted in 3.5 ml ice-cold PBS and were washed 3 times in 15 ml PBS by centrifugation at 1,350 x g for 5 min at 4 °C, to remove host cell RNA and debris.

The resulting pellets containing EP, SE/MT, ADT, and the three time-points for AMA (AMA₆, AMA₂₄ and AMA₁₂₀) were resuspended in ice-cold PBS and kept on ice. For bulk RNA-seq experiments, 2×10^7 parasites per sample were pelleted and frozen on dry ice until RNA extraction. For scRNA-seq experiments, parasites were diluted in PBS to 1×10^6 cells/ml.

15.1.3 Cell viability assessment

Aliquots from all parasite stages diluted at a density of 1×10^6 cells/ml were stained for 5 min at room temperature with SYTOX™ AADvanced™ Dead Cell Stain (Life Technologies, Eugene, OR, USA), following manufacturers’ recommendations. A total of 20,000 events per sample were acquired using a CytoFLEX flow cytometer (Beckman Coulter, Indianapolis, IN, USA) set to excitation at 488 nm and emission at 647 nm wavelengths. Percentage of viable cells was determined with the CytExpert v2.4 software (Beckman Coulter). Parasites killed with three cycles of freezing and thawing were used as

positive control for SYTOX AADvanced staining.

15.1.4 Bulk RNA-seq – RNA extraction, library preparation and sequencing

Total RNA was extracted using RNeasy Mini Kit (Qiagen, Hilden, Germany) following the manufacturer's protocol, with an on-column DNase digestion step with the RNase-Free DNase Set (Qiagen). Two biological replicates for each life cycle stage were prepared and sent to Novogene Co., Ltd (Cambridge, UK) for rRNA removal, library construction and sequencing. RNA quantity and integrity were measured with an Agilent 2100 bioanalyzer using the RNA Nano 6000 Assay Kit (Agilent Technologies, Santa Clara, CA, USA) and by agarose gel electrophoresis. A total of 1 µg total RNA per sample was used as input material for the lncRNA library preparation. rRNA was depleted from total RNA and strand-specific libraries were generated using NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (New England BioLabs, Ipswich, MA, USA) following manufacturer's recommendations. Index codes were added to attribute sequences to each sample.

Quality of the library was assessed by Agilent 2100 bioanalyzer, followed by quantification by qRT-PCR. The different libraries were pooled according to the effective concentration and the target amount of data – 6 Gb raw data per sample – and were sequenced by the Illumina NovaSeq 6000 (Illumina Inc., San Diego, CA, USA). Paired-end 150bp reads were generated.

15.1.5 Single-cell RNA-seq – library preparation and sequencing

Single-cell RNA-seq experiments were performed using Chromium Single Cell Gene Expression solution (10X Genomics, Pleasanton, CA), according to the manufacturer's guidelines. Single-cell suspensions of 16,500 cells in PBS were prepared for SE/MT and ADT (Fig.1B). AMA₆ was combined with AMA₂₄, and EP was combined with AMA₁₂₀ in equal proportion, in two cell suspensions of 16,500 cells each. Two technical replicates were collected for the SE/MT sample. A second biological replicate was prepared in a separate experiment by pooling equal quantities of EP, SE/MT, ADT, AMA₆, AMA₂₄ and AMA₁₂₀ for 21,000 cells in total. For this pool, two technical replicates were collected (labelled MIX1 and MIX2) (Fig.12B). Single-cell libraries were generated using the Chromium Next GEM Single Cell 3' Reagent Kit v3.1 and the Chromium Next GEM Single Cell 3' Reagent Kit v3.1 - Dual Index (10X Genomics) for the first and second biological replicates respectively. Briefly, single-cell suspensions were loaded onto a Chromium Controller instrument in a Chromium Next GEM Chip G (10X Genomics) to obtain single-cell Gel beads in EMulsion (GEMs). Barcoded cDNA was obtained from GEMs and amplified by PCR in a PCRmax Alpha Cyclor 1 thermal cyclor (Cole-Parmer, Saint Neots, UK). End repair and A-tailing, adapter ligation, post-ligation cleanup with SPRIselect (Beckman Coulter, Brea, CA, USA), and sample index PCR and cleanup steps were performed as per the manufacturer's instructions. Final library sample index PCR cycle parameters were selected based on quantification of barcoded cDNA samples with a Qubit 2.0 (Life technologies, Carlsbad, CA, USA), using the Qubit dsDNA HS Assay Kit (Life technologies, Eugene, OR, USA): SE/MT samples were amplified by PCR with 16 cycles while the rest of samples were amplified with 14 cycles. Single-cell libraries were sent to Novogene Co., Ltd for sequencing. Libraries were quantified by Qubit and quality was tested in an Agilent 2100 Bioanalyzer, before sequencing on the Illumina NovaSeq 6000 platform with PE150 strategy for 20 Gb raw data per sample.

15.2 Bioinformatics analysis

15.2.1 Creating reference transcriptomes for mapping

The 3' UTRs gene annotations of the *T. cruzi* Dm28c 2018 (Grisard et al. 2014) were extended by 2500 base pairs as described in Briggs and colleagues (Briggs, Rojas, et al. 2021). The Dm28c 2018 genome was appended with the *T. cruzi* maxicircle kDNA sequence. Two combined references were created with Cell Ranger v6.0.0. Both of the combined references contained the Dm28c 2018 3' UTR extended genome with kDNA maxicircle sequences, while the combined reference 2 also contained the GRCh38 human reference

15.2.2 Mapping and counting of life cycle bulk RNA-seq samples

All samples were mapped against combined reference 1 using STAR on default parameters (Dobin et al. 2013). Count matrices for the samples were generated using featureCounts (Liao, Smyth, and W. Shi 2014). Where multi-mapping and multi-overlapping genes were not counted, the minimum overlapping bases was 1 and paired reads were treated as a single fragment, not two individual reads. The MIX 1 dataset was also mapped against the Sylvio X10/1 genome (Franzen et al. 2011).

15.2.3 Processing and analysis of bulk RNA-seq samples

The genes whose read counts were less than 10 in 11 of our bulk RNA-seq samples were removed from the dataset, as well as the kDNA maxicircle genes.

To identify marker genes for the different life cycle stages captured, the following process was employed. Each life cycle stage was contrasted with another using the default parameters of the `results()` command in DESeq2 (Love, Huber, and Anders 2014). Genes were identified as markers of the life cycle stage/sample if they were significantly DE (i.e. they had a Benjamini-Hochberg adjusted p-value < 0.05) and higher expressed ($\text{Log}_2\text{FC} > 0.5$) in the life cycle stages/samples compared to the other life cycle stages/samples.

15.2.4 Mapping and counting scRNA-seq reads

For each sample, reads uniquely mapping to annotated genes on the created references were counted and assigned to a cell using the Cell Ranger Count function (Cell Ranger v6.0.0 - Default settings). Samples SE/MT and ADT were mapped against combined reference 2, while AMA_{6/24}, EP/AMA₁₂₀ and the MIX1 and MIX2 samples were mapped against combined reference 1.

15.2.5 Quality control and processing of scRNA-seq samples

Count data output from Cell Ranger was loaded into an AnnData object for analysis using Scanpy (Wolf, Angerer, and Theis 2018). Cell quality was assessed by looking at the total number of genes captured per cell, the percentage of reads mapping to the kDNA maxicircle, and the percentage of reads mapping to the human genome (for the samples that were mapped against combined reference 1, containing the human genome). Cells with high numbers of genes detected were removed as they could be doublets. Cells with low numbers of genes detected were removed as they may be dying cells or droplets which only contained ambient RNA. Cells with high percentage mapping to the kDNA maxicircle were removed as they may represent dying cells (Ilicic et al. 2016). Cells which had a high percentage of reads mapping

to the human genome were removed as we want to only capture individual parasite cells for analysis. The threshold values used for cell selection for each dataset are given in appendix table 1.

The count matrix of the AnnData (Virshup et al. 2021) objects for each sample were normalized through the following process. A cell's counts were divided by the total sum of counts for the cell, multiplied by 1000, before 1 was added to each value of the count matrix (to remove 0 values), and then finally the counts were transformed by natural logarithm. As some of the samples contain homogenous populations of cells (e.g. the trypomastigote sample, which just contains trypomastigotes), no filtering out of genes was performed, in order to retain genes which may uniquely define these subsets of these seemingly homogenous populations.

The individual AnnData objects were then concatenated to make a single AnnData object, and the normalised count matrix of all the genes and cells was stored in the raw layer of the AnnData object. The top 2500 most variable genes (in terms of normalized dispersion) were found for the concatenated dataset; these genes will be used for the scaling, dimensionality reduction, integration and clustering analyses outlined below. As the entire kDNA maxicircle was compressed down to two genes (forward and reverse sequence), the kDNA maxicircle genes were removed from the top 2500 most variable genes list in order to make sure the levels of kDNA gene expression was not driving the clustering or dimensionality reduction.

The effect of total number of genes per cell was regressed out of the count matrix in the active layer of the AnnData object, before the data was scaled to a mean of 0 and unit variance. Any scaled values above 10 were clipped to 10. PCA was performed on the scaled and regressed matrix, with the first 50 principal components (PCs) being calculated.

15.2.6 Integration of *T. cruzi* atlas samples

The Harmony (Korsunsky et al. 2019) and Seurat V5 CCA (Y. Hao, T. Stuart, et al. 2024) integrations were carried out in R (version 4.1.0). The mitochondrial genes were removed from the expression matrices. The expression values were then normalised using the NormalizeData function and the top 2500 variably expressed genes were identified. The normalised expression values were scaled and the PC embeddings of the cells identified. The PC embeddings were used for the basis for Harmony and Seurat integration. UMAP and Louvain clustering was then run on the first 10 integrated dimensions.

The final atlas integration was carried out using BBKNN (Polanski et al. 2020), with the batch key being the metadata column containing which sample each cell came from. Default values were used, apart from the number of PCs which was chosen to be 17, based on the variance ratio plot of the first 50 PCs. The BBKNN neighbour graph was used as the basis for creating a Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018) embeddings of the data, and assigning cells to clusters through the Leiden algorithm (Traag, Waltman, and Eck 2019) at a resolution of 0.7.

15.2.7 Cell cycle analysis of the *T. cruzi* atlas

The 305 *T. cruzi* cell cycle markers were taken from supplementary table 4 of Chávez and colleagues (Chávez, Eastman, et al. 2017). Cell cycle mapping was carried out as described in Briggs and colleagues (Briggs, Marques, et al. 2023). Briefly, for each individual dataset, the cell cycle genes were filtered down to those which had at least one transcript captured in 10% or more of the cells in the dataset. For each of the cell cycle phases (G1, S, G2), a score was calculated using the Seurat function MetaFeature. As the original code was written in R, it was converted into equivalent Python code so it could be used on AnnData objects. In order to express this score as a ratio, the MetaFeature score was divided by

the mean score for each phase. Cells were assigned the phase based on whichever phase had the highest score ratio, except if the cell had a phase ratio score of less than 1.05 for all phases, in which case it was assigned as ‘Unlabelled’. The same process was repeated using the PCF and BSF *T. brucei* cell cycle markers derived from Briggs and colleagues (Briggs, Marques, et al. 2023). The G1e and G1l genes were combined into one G1 gene list.

15.2.8 Analysis of epimastigote to metacyclic trypomastigote transitions in scRNA-seq data

The epi_2, meta_3 and epi_meta_trans 7 and 8 clusters were extracted from the atlas object into their own object. All subsequent analysis was performed on this subsetted dataset.

The data was rescaled and reclustered in Scanpy before ForceAtlas (FA) 2 (Jacomy et al. 2014) embeddings were created for the cells and pseudotime analysis carried out with PAGA (Wolf, Hamey, et al. 2019). Cells with PAGA pseudotime values more than 0.65 were removed from the dataset. The data was then converted to SingleCellExperiment (Amezquita et al. 2020) format for fitting of the general additive models using tradeSeq (Van den Berge et al. 2020), across the PAGA pseudotime. For both datasets, the models were fitted over 7 knots. An association test was then carried out to identify which genes have significant expression changes over the pseudotime axis, where significant genes were those with a \log_2FC more than 0.25, a Bonferroni corrected p-value of less than 0.05 and expression in at least 10% of cells in the dataset.

To generate the smoothed expression plot sorted by expression peak, the expression values of the genes identified by tradeSeq as being DE were smoothed using the predictSmooth() function. The smoothed expression values were then scaled, before they were fourier transformed using the FitWave function from the metR library. The genes were then ordered by peak across pseudotime. Genes were defined as early, middle or late based on the following criteria: Smoothed expression pseudotime between 0-16 = early, 17 - 33 = middle, 34-50 = late.

15.2.9 Analysis of trypomastigote subsets

The trypomast_0 and trypomast_4 clusters were subsetted into their own object and the significantly DE genes (absolute $\log_2FC > 0.5$, Bonferroni corrected p-value < 0.05) between the two clusters were identified.

15.2.10 Reference mapping using the *T. cruzi* atlas

The following training to test data proportion sets were created: 0.95:0.05, 0.75:0.25, 0.5:0.5, 0.25:0.75 & 0.05:0.95. The training sets were then used to train a scPoli (De Donno et al. 2023) model and predict the cell type of the test dataset. This process was repeated once more before weighted average f1-scores were calculated. ScPoli was run using mainly default parameters apart from: an embedding dimension of 7, the condition keys being the sequencing run, and sample IDs when training on the atlas training set.

15.2.11 Annotating an unrelated scRNA-seq run using the *T. cruzi* atlas

A new scRNA-seq dataset was generated from amastigotes isolated from 48 hpi cultures. Sample collection and library preparation followed the procedures described in sections 2.1.2, 2.1.3 and 2.1.5, except

that the cultures of Vero infected cells were washed at 16 hpi rather than 6 hpi. A scPoli model was then trained on the entire *T. cruzi* atlas and used to predict the cell type label of the 48 hpi amastigotes.

15.2.12 Package versions

STAR version 2.7.11a

FeatureCounts version 2.0.6

Analyses in Python were carried out on Python version 3.8.16. Specific package versions for Python analysis are as follows: anndata – 0.9.1, h5py – 3.8.0, leidenalg – 0.10.1, matplotlib – 3.7.1, numpy – 1.24.3, pandas – 2.0.3, pip – 23.0.1, scanpy – 1.9.3, scvi-tools – 0.20.3, scArches – 0.5.10, seaborn – 0.12.2, scipy – 1.19.1, scikit-learn – 1.2.2, umap-learn – 0.5.3, bbknn – 1.6.0, fa2 – 0.3.5, igraph – 0.10.6, pytorch-lightning – 1.9.5

Analyses in R were carried out on R version 4.1.0. Specific package versions for R analysis are as follows: singlecellexperiment – 1.14.1, tradeSeq – 1.6.0, metR – 0.15.0, ComplexHeatmap – 2.8.0

16 Chapter 2: Discussion

In this chapter I have shown that not only can the transcriptome of *T. cruzi* be captured using scRNA-seq, but that the information captured can allow a finer resolution view of the life cycle stages of *T. cruzi* and the transition between epimastigote and metacyclic trypomastigote cells.

In the atlas, we defined clusters for all four of the major life cycle stages of *T. cruzi*, transitional cells between epimastigotes and metacyclic trypomastigotes, different development stages of amastigotes and trypomastigote subsets, and validated their identity through matched bulk RNA-seq data and established literature.

When predicting the life cycle stage annotations, the f1-scores varied between the life cycle stages. The lowest scores tended to be the smaller clusters, such as the `epi_meta_trans_7` and `trypomast_9`, and thus their low scores could be explained by the lack of adequate training data for the task. Furthermore, the weighted average f1-score tended to be ~ 0.75 across the different training subsets. This score fits into the ranges of weighted average f1-scores scPoli achieves for human datasets (De Donno et al. 2023), meaning the predictive ability of the *T. cruzi* atlas is on-par with human datasets. With the annotation of the 48 hpi amastigotes, we show that the atlas can be used to annotate novel data which is not represented in the atlas. This result is especially important as one of the primary uses of the atlas is to allow researchers to perform preliminary annotations of their own data using the atlas.

ScRNA-seq data from *T. brucei* (Briggs, Marques, et al. 2023) and *Leishmania major* (*L. major*) (Marques, Laidlaw, Briggs et al, unpublished data) show that cells can be separated out and annotated based on their cell cycle phase using scRNA-seq data. The lack of ability to replicate this in *T. cruzi* using either a list of *T. cruzi* derived genes (Chávez, Eastman, et al. 2017) or syntenic orthologs from *T. brucei* is thus an interesting observation. One reason for why this could be the case is that Chávez and colleagues derived the cell cycle markers using DESeq (Anders and Huber 2010), which has been obsolete since the publication of DESeq2 in 2014 (Love, Huber, and Anders 2014). Furthermore, when

testing significance of a gene, they did not compare one cell cycle phase against all the others, but rather which ones were next in the cell cycle. For example, to derive the cell cycle markers of S phase, they compared the S phase sample with the G2M sample. On top of this, they only have one sample per cell cycle phase, meaning DESeq2 cannot be run as it requires at least two samples per condition tested. Given the low number of samples and out-of-date analysis software, this could lead to false positive cell cycle associated genes being present in the list, leading to inaccurate cell cycle phase labelling. While this could explain why the *T. cruzi* derived genes did not lead to successful annotations, the cell cycle phase annotations for the *L. major* analysis were derived from syntenic orthologs of the *T. brucei* genes and were accurate, which is not the case for the *T. cruzi* data.

An alternative explanation for why the cell could not be correctly annotated is that cell cycle is regulated at the transcriptome or proteome level. This would fit with the findings of this thesis chapter, where the *T. cruzi* cells could not be successfully annotated with a list of cell cycle phase marker genes derived from *T. brucei* or *T. cruzi* itself. In fact, another paper by Chávez and colleagues (Chávez, Urbaniak, et al. 2021) suggests that the parasite does regulate cell cycle at the levels of transcriptome and proteome, although why this differs from *Leishmania* and *T. brucei* remains an interesting question.

The other possible option is that the *T. cruzi* cell cycle phase is able to be accurately annotated using transcript level data, but the genes involved are not orthologous to ones in *T. brucei*, nor were they correctly derived in the paper by Chávez and colleagues (Chávez, Eastman, et al. 2017). To confirm whether or not this is the case, a new experiment should be carried out where multiple bulk RNA-seq samples are collected for each phase, and DESeq2 is used to analyse the data.

Trypomastigotes in the blood and in culture have two morphological forms: slender and broad. Slender forms display increased infectiveness of host cells in comparison to the broad forms as they can infect cells via direct invasion of host cells, while the broad forms infect through phagocytosis (Brenner and Chiari 1963, Schmatz, Boltz, and Murray 1983 & Martin-Escolano et al. 2022). Other research has demonstrated that there exists two subsets of trypomastigote, which have high and low expression of trans-sialidase protein, with the trans-sialidase high cells being able to effectively invade host cells, while their lowly expressing counterparts having significantly reduced infection capability (Pereira et al. 1996). While the previously cited paper did not comment on the morphology of the two subsets, and the images taken of the two subsets are not high definition enough to distinguish morphological differences, these could represent the broad and slender form trypomastigotes. In our data we have captured two clusters, one with high trans-sialidase (trypomast_4) and one with low trans-sialidase expression (trypomast_0), which may correspond to these two subsets.

Having characterised functional differences between the two subsets, the next question is what factors might influence the cells to go towards these different forms. As stated above, RBPs can influence cellular processes and differentiation, and thus may influence trypomastigote form. Orthologs of the *T. brucei* genes RBP10 and ZC3H32 were markers of the trypomast_4 cluster, both of which are associated with the BSF stages of *T. brucei*. ZC3H32 knockout is fatal to BSF but not PCF (C. Klein, Terrao, and Clayton 2017), and RBP10 is only expressed at the protein level in BSF forms (B. Liu and Clayton 2022). Furthermore, lack of RBP10 expression in BSF *T. brucei* causes them to transition into PCFs, and expression of RBP10 in PCFs causes them to transition to BSFs. These previous findings mark these RBPs as being key markers of parasites while in the mammalian host and, combined with the findings of this analysis, may implicate them as a marker of invasive trypomastigotes.

Ultimately, these findings are speculative as there is uncertainty over whether these two clusters represent the two trypomastigote subsets, and no wetlab validation has been carried out of these findings.

To date, only cell morphology, infection capability and, perhaps, trans-sialidase protein expression have been used to distinguish the subsets. Future research should thus be undertaken to isolate these two populations, and to perform transcriptomic and proteomic analysis on the samples to see whether or not these two clusters match to the biological relevant subsets of trypomastigote.

Through the interpretation of the genes that were found to be significantly changing in expression across the pseudotime axis, we show that the process of metacyclogenesis has been captured. The downregulation of L-threonine metabolism genes over the course of metacyclogenesis may reflect the parasite preadapting to the mammalian environment by switching off redundant metabolism associated processes, similar to the metacyclic forms in *T. brucei* (Christiano et al. 2017). The upregulation of Chagasin towards the end of the pseudotime axis demonstrates that the metacyclic trypomastigotes may be preparing to invade host cells when passed back into the mammalian host.

The finding that several genes encoding homologous recombination proteins peak during the middle of metacyclogenesis is an interesting finding. Previous studies have shown that overexpression of RAD51 in epimastigotes promotes resistance to some oxidative stresses, and overexpression in amastigotes leads to higher intracellular parasitaemia (Gomes Passos Silva et al. 2018). The roles of the recombination proteins in metacyclogenesis has not been commented on. Interestingly, mRNA levels of RAD51 positively correlate with dormancy in epimastigotes, and heterozygous knockout of RAD51 in CL Brener parasites leads to a reduction in dormant epimastigotes (Resende et al. 2020). Dormancy as a biological concept has been discussed in *T. cruzi* and bacteria and fungi (Harms, Maisonneuve, and Gerdes 2016 & Sánchez-Valdéz et al. 2018) with the common definition being that dormancy represents a reversible state a cell enters where its growth and metabolism slow down, allowing it to survive certain stresses. Exposure to drugs are one such stress, with dormant *T. cruzi* amastigotes generally more resistant to drugs than non-dormant cells (Sánchez-Valdéz et al. 2018). Interestingly, these dormant amastigotes occur without initial drug stresses, suggesting they are not induced into dormancy by drug exposure; leaving the driver of dormancy in amastigotes unknown (Sánchez-Valdéz et al. 2018). In the paper by Resende and colleagues (Resende et al. 2020), they induce epimastigote dormancy through starving the cells of nutrients over six days, leading to the generation of stationary epimastigotes. Whether or not stationary epimastigotes are dormant cells is up for debate. Stationary epimastigotes are cell cycle arrested and undergo metabolic changes in comparison to exponential phase epimastigotes (Barisón et al. 2017), which could fit the definition of cellular dormancy. As nutrient starvation is used to induce metacyclogenesis (MacLean et al. 2018), factors which influence cell dormancy may be important in the process, such as the recombination proteins.

Along with metabolic, repair and invasion protein gene expression changes, we also identify changes in expression of RBP genes across the pseudotime axis that are known to be associated with metacyclogenesis. This gives further evidence that this process is captured in the atlas data.

While the markers of metacyclogenesis mentioned previously validate that we have captured the process of metacyclogenesis in the atlas data, they do not reveal insights into the process that were not previously described. It is thus important that we have not only captured known markers of metacyclogenesis, but also identified RBPs which have not previously been associated with metacyclogenesis. These markers thus represent possible novel findings in terms of metacyclogenesis associated markers, and work should be done to generate KO models for these genes to see whether the KOs affect the process of metacyclogenesis.

With this analysis we have established a bedrock of *T. cruzi* transcriptomic heterogeneity, which future researchers can build their research upon. That is not to say we have created an exhaustive atlas

of the parasite transcriptome. The main transition captured in the atlas was between the epimastigotes and metacyclic trypomastigotes, leaving the trypomastigote to epimastigote and trypomastigote to amastigote transitions uncaptured due to them not being carried out and the amastigote to trypomastigote transition uncaptured due to large differences in transcriptome. Furthermore, these transcriptomic changes are reflective of what occurs in parasite under *in vitro* conditions, which may not reflect what changes occurs in the parasite *in vivo*.

All of the above is not to say that this atlas is a perfect representation of the *T. cruzi* life cycle, as certain observations exist which do not fit with the biology of the system. One example of this is that while the AMA_{6/24} sample has many cells present in the early_mid_amast_5 cluster, these cells also make up part of the epi_meta_trans_7 cluster. There are a variety of reasons why this could have occurred, the most likely of which are due to difficulties with integration and difficulties working with non-model organisms. We tested out Seurat V5 CCA, BBKNN and Harmony to integrate the datasets, and came to the conclusion that BBKNN and Harmony performed well, but that doesn't mean they performed perfectly, and some cells could be badly integrated. Integration of parasite data is itself a problematic procedure as the tools used to integrate were most likely only tested on mammalian datasets which, in general, have better annotated transcriptomes and thus more dynamic expression data. This becomes increasingly strained for Trypanosomatids which are not only parasites, but have a different mode of gene expression to other parasites and mammals. As stated in the introduction, the combined problem of many integration methods, but lack of consistent ways to assess integration method performance, makes choosing the correct integration method for datasets difficult. On top of this, most trypanosomatid scRNA-seq datasets are of *T. brucei* (Briggs, Rojas, et al. 2021, Briggs, Marques, et al. 2023, Vigneron et al. 2020, Howick, L. Peacock, et al. 2022 & Hutchinson et al. 2021), the most studied trypanosomatid. The presence of AMA_{6/24} sample cells in the epi_meta_trans_7 cluster may also explain why it is the second most predicted cluster ID for the 48 hpi amastigote dataset analysis.

Although we have presented biological evidence for our life cycle stage assignments to clusters, we cannot be 100% confident that the clusters represent what we say they do, and thus future users of the data should be wary of this limitation. Despite limitations, this atlas is the first step in dissecting the *T. cruzi* life cycle at single cell resolution to better understand what transcriptomic changes define the transitions between its host and vector stages and the variability of cell phenotype within specific life cycle stages.

17 Chapter 3: Transcriptomic analysis of *RBP6* overexpression induced *T. brucei* insect life cycle stages

Unlike the mammalian stages (and their transitions) which can be captured *in vitro*, the insect stages of *T. brucei* (under WT conditions) requires the presence of its vector: the tsetse fly. This makes capturing the insect life cycle stages more difficult and as a consequence knowledge of their transitions is not as great compared with the mammalian stages. In this chapter I implement the *RBP6* overexpression model of *T. brucei* insect stage development and apply scRNA-seq to the model in order to identify the drivers of life cycle stage transition. The results show that the transcriptomic signature of epimastigotes are seemingly absent in the *RBP6* overexpression dataset, while a population of PAG expressing cells is present which may be generated by the *RBP6* overexpression process. I also highlight a lack of consistency in integration method performance for *T. brucei* scRNA-seq datasets.

18 Chapter 3: Introduction

Within the insect and mammalian forms of *T. brucei* (*T. brucei*), the parasite undergoes many transitions between different distinct life cycle phases (Matthews 2005). The transitions within the mammalian stages, slender to stumpy (Reuner et al. 1997, Rojas et al. 2019), and between mammalian and insect stages, stumpy to PCF (Czichos, Nonnengaesser, and Overrath 1986, Dean et al. 2009, Schuster et al. 2021 & Ngoune et al. 2024), are well understood in terms of the mechanisms that allow their progression. This stands in contrast to that of the insect stage life cycle transitions, where the mechanisms that drive the transitions are largely unknown. While the slender to stumpy and stumpy to PCF transitions for wild type cells can be captured *in vitro* (Cross and Manning 1973, H. Hirumi, Doyle, and K. Hirumi 1977, Czichos, Nonnengaesser, and Overrath 1986), this is not the case for the transitions in the insect stages, i.e. PCF to epimastigote and epimastigote to metacyclic. Like many organisms, control of transition in *T. brucei* is partially done through changes in transcriptome. As such, bulk and single cell RNA-seq techniques have been applied to shed more light on what drives the transitions in these stages, revealing gene expression differences in metabolism, surface coat, environmental responses, among others (Telleria et al. 2014, Vigneron et al. 2020, Hutchinson et al. 2021, Howick, L. Peacock, et al. 2022). *In vivo* work is required to capture these transitions by infecting tsetse flies with stumpy form parasites. However, there are complications with this approach such as the efficiency of infection (Ngoune et al. 2024), and the need to isolate the parasites from the tsetse fly itself. Thus, ways to try and capture the insect life cycle stages *in vitro* has been of key interest to *T. brucei* researchers.

The polycistronic nature of gene expression in the parasite means that it is largely facilitated through controlling mRNA stability and splicing post transcriptionally, rather than through controlled activation of RNA II polymerase activity (Fernandez-Moya and Estevez 2010). Therefore, mRNA binding factors are important in facilitating the transitions in the life cycle forms. Because of their role in gene regulation, targeting these mRNA binding factors has been shown to cause changes in the steady state of cells. The first paper to describe such an effect was from Kolev and colleagues (Kolev, Ramey-Butler, et al. 2012), who showed that overexpressing the gene encoding RNA binding protein 6 (*RBP6*), in PCFs over several days, causes the emergence of morphologically epimastigote and metacyclic stage cells, and the switching on of VSG protein expression. The metacyclics generated using the *RBP6* overexpression model can be transitioned to mammalian stages through *in vivo* infection of mice, but not *in vitro* (Fig.18). *In vitro* transitions of *RBP6* metacyclics to mammalian stage forms would be shown in a later paper by the authors in which they isolated a mutant strain of *T. brucei* where a single point mutation (substituting a glutamine (Q106) for a lysine) in the *RBP6* gene allowed for the *in vitro* transition of metacyclics into the mammalian stages when *RBP6* was overexpressed (H. Shi, K. Butler, and Tschudi 2018). The only downside to this second model of *T. brucei* development, is the apparent lack of epimastigote generation, as shown by the lack of brucei alanine-rich protein (BARP) protein expression (Fig.18). Further papers highlighted more factors involved in differentiation, most notably RBP10, where RNA interference (RNAi) caused slender stage parasites to become cell cycle arrested, express the stumpy specific marker PAD1, and be able to transition into PCFs when RNAi was induced (Mugo and Clayton 2017) (Fig.18). Furthermore, the authors also showed that expression of RBP10 in *in vitro* cultured PCFs (derived from mammalian form cells) had reduced growth, but proliferated and expressed variable surface glycoprotein (VSG) when transferred to mammalian form culture conditions. Whether these cells entered a stumpy stage is unknown (Mugo and Clayton 2017) (Fig.18).

As some insect form life cycle stages are left out in the *RBP10* and *RBP6* point mutation models, this leaves the *RBP6* overexpression model the *in vitro* approach which appears to most accurately reflect the *in vivo* life cycle of the parasite. The model has already been used to examine metabolic changes that occur across the insect stage transitions (Dolezelova et al. 2020), so its applicability to investigate the different insect stages has been validated, and thus would make a good model for inves-

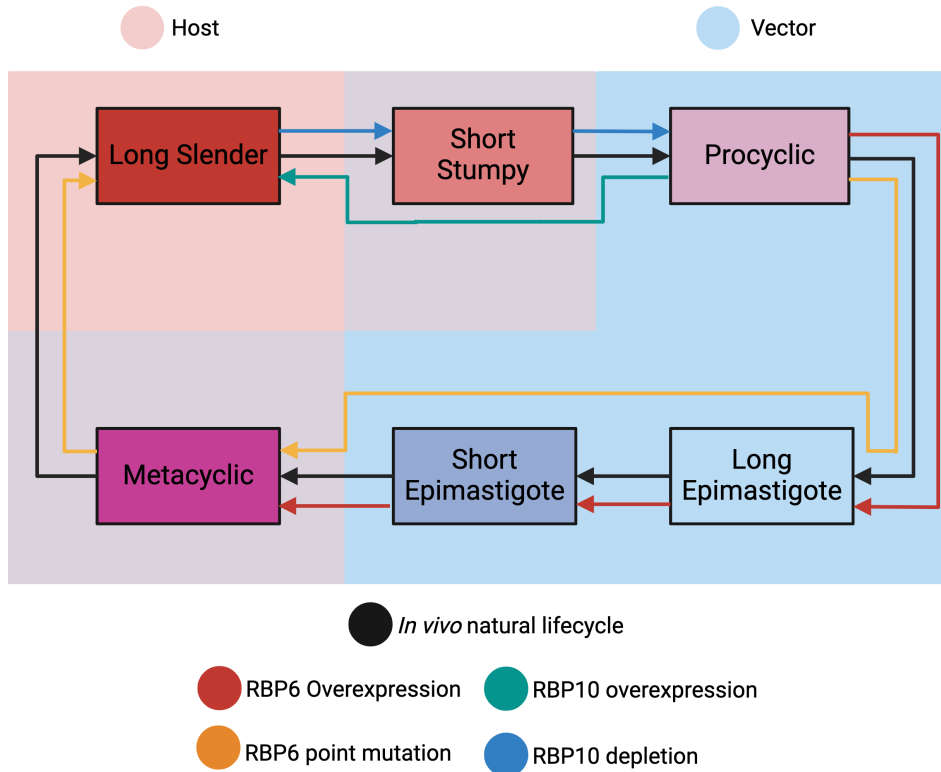


Figure 18: **Diagram showing the progression of the *in vivo* *T. brucei* life cycle and genetic alteration mediated *in vitro* life cycles**

Diagram showing how the *T. brucei* life cycle progresses across the *in vivo* life cycle (black line) and various *in vitro* models of the life cycle such as: *RBP6* overexpression (red), *RBP6* mutant overexpression (orange), *RBP10* overexpression (green) and *RBP10* deletion (blue). The background of the diagram details whether the life cycle stages are found in the host (pink) or vector (light blue). A mix of the two colours shows that the life cycle stages are found in both.

Investigating the transcriptomic changes that occur during transitions between the insect stage forms. As cells do not necessarily transition in synchrony, methods like bulk RNA-sequencing are ill advised to capture the transcriptome of asynchronous cell-cell transitions, as their transcriptomic signal could be silenced or clouded by the signal from non-transitional cells. Single cell RNA-sequencing (scRNA-seq), on the other hand, allows the transcriptome of individual cells to be profiled, avoiding the issue of clouded signal. In the context of *T. brucei*, scRNA-seq has been applied across various papers to profile different life cycle stages of the parasite and its transitions. The first application was by Müller and colleagues (Müller et al. 2018) who used scRNA-seq to show that mVSG gene expression switching can be induced through deletion of histone variants. Vigneron and colleagues (Vigneron et al. 2020) were next, profiling the SG of infected tsetse flies to better understand the process of metacyclogenesis. Through this, the authors identified several surface marker genes of metacyclic cells, notably *SGM1.7*. Briggs and colleagues (Briggs, Rojas, et al. 2021) were the first to analyse the mammalian stages of the parasite, capturing the transition of slender into stumpy BSFs under WT conditions and when the transition is blocked due to knockout of the gene encoding *ZC3H20*. Hutchinson and colleagues (Hutchinson et al. 2021) captured cells from the SG of the tsetse fly in order to identify the dynamics of mVSG expression, identifying pre-metacyclic cells which transcribe more than one mVSG. Howick and colleagues (Howick, L. Peacock, et al. 2022) also utilised scRNA-seq to analyse cells in the SG, focusing on the sexual forms of the parasite and identifying conserved markers of these forms.

In this chapter, we implement the *RBP6* overexpression in *T. brucei* PCFs *in vitro*, and confirm that it leads to the emergence of morphologically epimastigote form cells. I also perform scRNA-seq on the *RBP6* overexpression PCFs at three pooled timepoints post inducement. I show that the scRNA-seq and previously described bulk RNA-seq results mirror each other in terms of transcriptional changes of marker gene expression at different points post inducement. I also show an apparent lack of epimastigote cell generation by the *RBP6* overexpression system as well as the possible presence of life cycle stages which may not have a parallel in the *in vivo* system. Finally, we show that the transcriptome of *RBP6* overexpression cells matches neither *in vivo* derived insect stage cells or *in vitro* non-differentiating PCFs, with many genes differentially expressed between the datasets, including key metabolic genes.

19 Chapter 3: Results

19.1 Implementing *RBP6* overexpressing PCFs

Primers specific to the 427 Lister strain *RBP6* gene (Tb427.03.2930) were taken from the original authors paper (Kolev, Ramey-Butler, et al. 2012), with a one base substitution in the reverse primer (to match the genomic sequence of the 427 Lister *RBP6* gene), and used to PCR-amplify the *RBP6* nucleotide sequence from 427 Lister gDNA. The PCR product was analysed using gel electrophoresis, a band between the 600 and 800 base pair (bp) DNA ladder bands was seen (Fig.19A). The predicted sequence length of the *RBP6* Open Reading Frame (ORF) in Lister 427 is 726 bp, which would fit with the location of the band within the gel; thus a sequence of DNA from the Lister 427 gDNA of roughly the same length as the *RBP6* gene in Lister 427 was successfully amplified.

To engineer a plasmid which would allow the overexpression of the *RBP6* gene, the pLEW100v5 plasmid was utilised (Wirtz et al. 1999), as it allows for tetracycline-induced overexpression of a gene in parasites which have both a tetracycline repressor and the T7 RNA polymerase. Digestion of the plasmid with the restriction enzymes BamHI and HindIII should separate the plasmid into the insert (1877bp in length) and the backbone (5808bp in length). The digested plasmid, ran on an agarose gel,

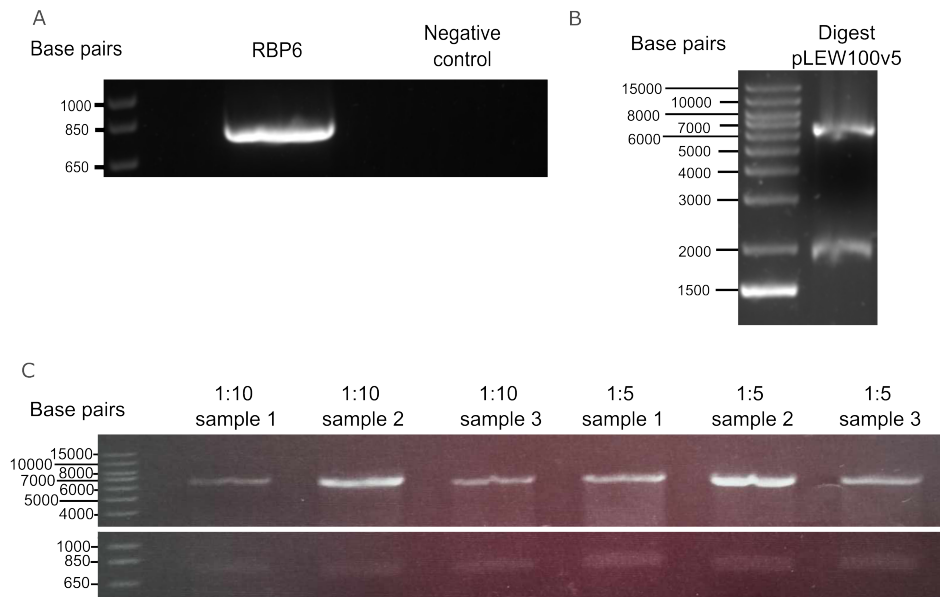


Figure 19: Gel electrophoresis results for the amplification of *RBP6*, digestion of pLEW100v5 and digestion of engineered *RBP6* overexpression plasmid

Photographs of agarose gels from gel electrophoresis experiments of the digested pLEW100v5 plasmid (A), PCR amplified *RBP6* gene + negative control without polymerase (B) and the cut engineered plasmid containing the pLEW100v5 backbone and the *RBP6* gene insert (C).

showed two strong bands near the 2000bp and 6000bp bands of the DNA ladder (Fig.19B). These results show that the plasmid can be successfully separated into the backbone and insert.

The backbone of the plasmid was extracted from the gel and purified prior to ligation with the amplified *RBP6* ORF, creating the *RBP6* overexpression plasmid (OE-plasmid). To see if the length of the new insert matched that of the *RBP6* gene, a digestion with HindIII and BamHI was carried out. The results show two bands of similar lengths to the 850bp and 6000bp DNA ladder bands (Fig.19C), showing that the insert of the OE-plasmid has the same bp length as the *RBP6* gene. To ensure that the *RBP6* gene sequence was successfully ligated into the OE-plasmid without loss of information, Sanger sequencing was performed on the insert site of the OE-plasmid, utilising primers which bound upstream (primer JM42) and downstream (primer HDK430) of the insertion site. The sequences were aligned to the Lister 427 *RBP6* ORF using EMBOSS Needle (Madeira et al. 2024) (appendix file 19 & 20) as well as Benchling, with both of the primers binding the *RBP6* ORF, however the HDK430 primer bound on the *RBP6* ORF itself, rather than downstream of it (Fig.20A). The alignments revealed possible frameshifts and single nucleotide polymorphisms, although many of these were not captured by both the forward and reverse Sanger sequences (Fig.20C). Both sequences however showed a deletion of a CAA repeat compared with the reference *RBP6* ORF (Fig.20B).

The OE-plasmid was then cloned into transformation competent DH5- α *E. coli* to allow OE-plasmid stocks to be replenished if the original batch was depleted. Several transformation attempts were made, leading to the generation of a plate with clones, which were isolated and stored long-term at -80°C. The linearized OE-plasmid was then transfected into the tetracycline-induced overexpression competent 29:13 Lister 427 PCFs (Wirtz et al. 1999), with two clones generated after drug selection: clone C1 and clone

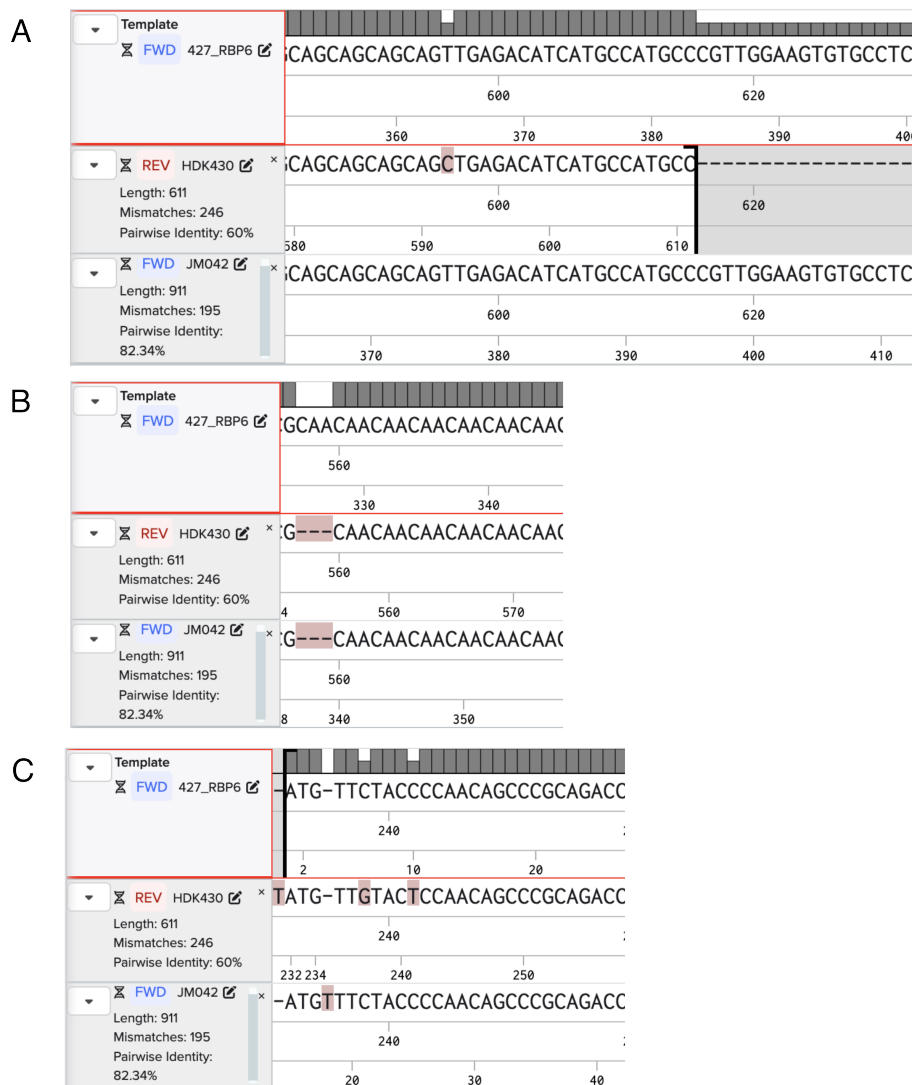


Figure 20: Alignments of the Sanger sequences against the 427 Lister *RBP6* ORF

Various screenshots showing the nucleotide alignment of the forward (bottom track) and reverse (middle track) Sanger sequences against the *RBP6* ORF from Lister 427 (top track). These visualisations show the start of the reverse Sanger sequence (A) as well as deletions (B) and frameshift and single nucleotide polymorphisms (C) in the Sanger sequences compared to the *RBP6* ORF.

D5.

One of the effects of *RBP6* overexpression induction should be a decrease in growth of the parasite (Kolev, Ramey-Butler, et al. 2012). This is due to the fact that metacyclics are quiescent, unlike the PCFs and epimastigotes. Thus, when *RBP6* overexpression is induced in the transfected clones, there should be a decline in the growth curve. The C1 and D5 clones were split into two cultures each, one where *RBP6* overexpression was induced with 10 $\mu\text{g}/\text{ml}$ of tetracycline and one where it was not. The parasites were then cultured for seven days, with the parasites being passaged down to 2×10^5 cells/ml each day. While the uninduced clones maintained approximately the same density across the seven days,

a steady decline in density of the induced clones can be seen over the same time period (Fig.21). These results provide evidence that *RBP6* overexpression is working, in that, generation of non-replicating cells by *RBP6* overexpression would reduce cell density relative to the uninduced at each passage, but does not prove it.

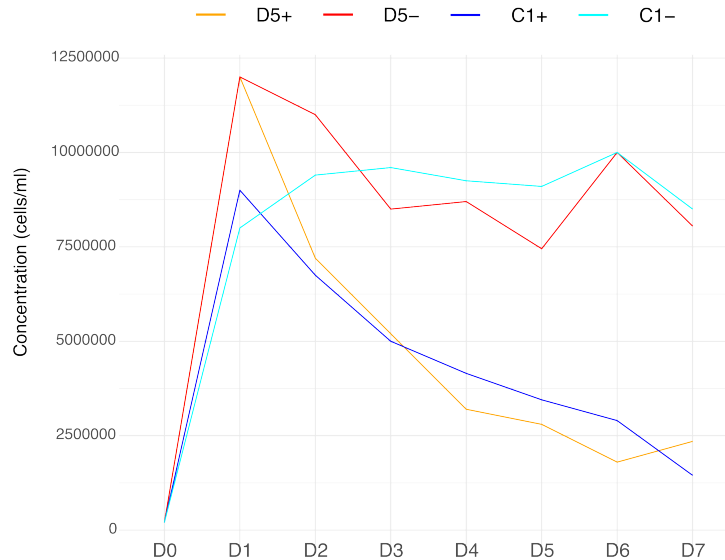


Figure 21: **Growth curve of the engineered *RBP6* overexpressing clones**

Concentrations of the C1 and D5 *RBP6* overexpression clones when induced to express *RBP6* over seven days. For each clone, an uninduced culture was kept for the same length of time. The cells were cultured in SDM-79. Each line represents one culture. The culture samples with '+' represent that they are induced with tetracycline while '-' indicates they were not induced by tetracycline.

As mentioned, Shi and colleagues (H. Shi, K. Butler, and Tschudi 2018) reported that a mutation in the *RBP6* gene, which causes a glutamine (Q106) to be replaced by a lysine, leads to major changes in the *RBP6* overexpression process. As one mutation can have such a major effect on the outcome of the *RBP6* overexpression process, the presence of many amino acid changes in the *RBP6* insert ORF of the engineered plasmid may complicate the results. Given these concerns, all future sections in this thesis which involve the *RBP6* overexpression model, were carried out using an established *RBP6* overexpression cell line gifted by Dr Lucy Glover of the Institut Pasteur. Furthermore, at the suggestion of Dr Lucy Glover, when the cells were being induced to overexpress *RBP6*, they were maintained in SDM-80, not SDM-79. SDM-80 and SDM-79 are very similar culture mediums, with SDM-79 containing a higher concentration of glucose than SDM-80 does. SDM-80 also contains a higher concentration of proline, thus SDM-80 is more reflective of the parasite environment (Schönenberger and Brun 1979 & Lamour et al. 2005).

To ensure the new *RBP6* overexpression parasites behaved as expected, a parasite growth curve of the cells induced to overexpress *RBP6* was compared with uninduced parasites. A growth defect in the *RBP6* overexpression induced parasites can be seen from day three onwards, however a 42.3% decline in uninduced sample cell density was seen between day three and day four, before increasing by 53.2% between day four and day five (Fig.22).

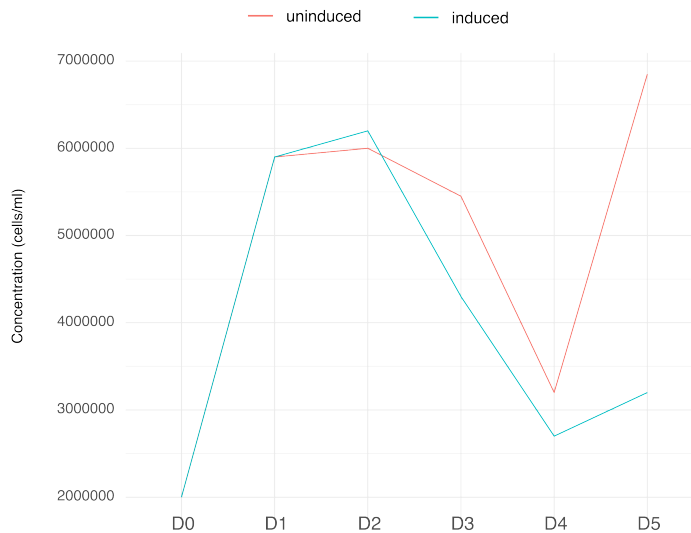


Figure 22: **Growth curves of the gifted *RBP6* overexpression cells**

Concentrations of the *RBP6* overexpression cells (gifted by Dr Lucy Glover) over five days of *RBP6* overexpression induction, as well as an uninduced control. The induced cells were cultured in SDM-80 while the uninduced cells were cultured in SDM-79. Each line represents one culture.

19.2 Image analysis of the Pasteur *RBP6* overexpressing cells

To validate that the *RBP6* overexpressing cells were changing into epimastigotes and metacyclics, an immunofluorescence assay (IFA) was carried out. To distinguish the different stages, an anti-EP procyclin and anti cross-reacting determinant (CRD) (the latter gifted to us by Dr Lucy Glover) was utilised. CRD is an epitope on the glycosyl-phosphatidylinositol anchor of the VSGs, meaning it can be used to distinguish metacyclics, which express VSG, from epimastigotes and PCFs, which do not (Zamze et al. 1988). EP procyclin, on the other hand, is expressed on PCFs and epimastigotes, but not on metacyclics (Acosta-Serrano et al. 2001). To separate PCFs from epimastigotes, the positioning of the DAPI-stained nucleus and kinetoplastid was assessed, as the kinetoplastid is closer to the posterior end of the cell than the nucleus in epimastigotes, while the opposite is true of PCFs (Sunter and Gull 2016).

IFAs were carried out for three, four and five days post- *RBP6* overexpression induction. Uninduced cells were also cultured for this length of time as a control. Information from Dr Lucy Glover suggested that these days were the optimal time points, as the cells should be at the boundaries of PCF to epimastigote, and epimastigote to metacyclic, transitions. The growth curve of the cells captured for IFA can be seen in Figure 22.

For each of the timepoints, IFAs of tetracycline induced and uninduced cells were taken. The induced and uninduced day four and induced day three samples were not processed due to bad slide quality and the day five uninduced sample slide was accidentally broken.

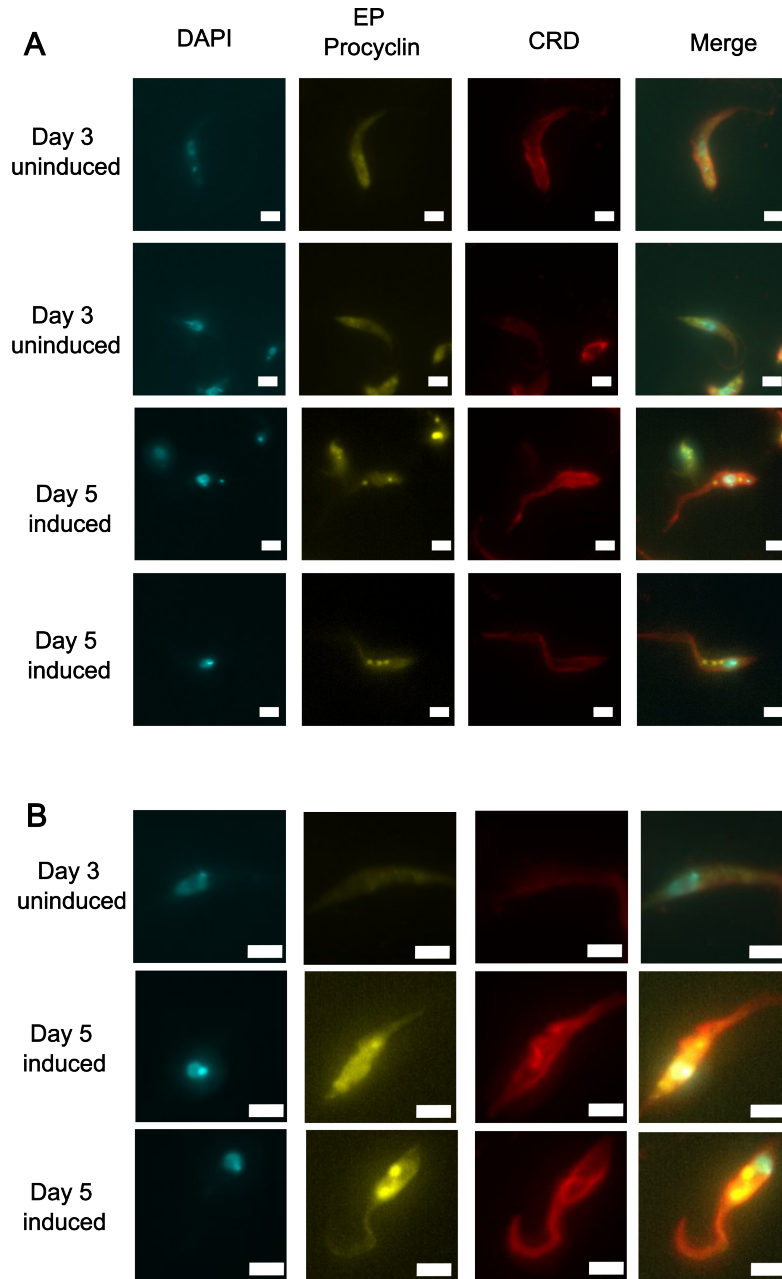


Figure 23: CRD and EP procyclin protein expression

IFA images of cells from uninduced cells three days post experiment start and *RBP6* overexpression induced cells five days post inducement (A). Representative images at day five post *RBP6* overexpression of possible epimastigote cells (B). Scale bars are $3\mu\text{m}$.

All cells in the uninduced day three sample were anti-EP procyclin and anti-CRD positive across the cell body, with some cells having visually higher fluorescence of anti-CRD than others (Fig.23A). In the uninduced day three cells, no anti-CRD fluorescence should be detected, as the cells should be PCFs. The presence of staining with anti-CRD suggests perhaps the antibody is non-specific or the overexpression of *RBP6* is leaky and thus metacyclic forms are generated. No cells with its kinetoplastid closer to the posterior of the cell than the nucleus could be seen in the day three uninduced sample, suggesting no epimastigote cells are present.

The induced day five sample cells were also anti-EP procyclin and anti-CRD positive, with some cells having visually higher fluorescence of anti-CRD than others (Fig.23A). The anti-CRD fluorescence was distributed across the cell body in all the induced day five cells, while the anti-EP procyclin fluorescence pattern was not distributed across the cell body in all the cells. Five cells (out of 19) had anti-EP fluorescence across the cell body (Fig.23B Top & Middle row), while 14 cells (out of 19) had localised expression of anti-EP that formed spots (Fig.23A). The anti-EP procyclin localisation on the day five post-inducement bears a striking resemblance to intracellular IFA images of PCFs, where EP procyclin is inhibited from being trafficked to the surface (L. Liu et al. 2013). These anti-EP spots could represent cellular debris, however the bright field view of the slide was obscured by general debris across all of the IFAs taken, making this possibility difficult to assess.

For the induced day five sample, cells with their kinetoplastid closer to the posterior end could be seen (Fig.23B Top row), suggesting epimastigotes were generated. Definitively identifying epimastigotes was difficult as many cells could be seen in which their kinetoplastid and nucleus overlapped (Fig.23B Middle & Bottom row).

These results of the IFA show that despite the fact that the uninduced sample should be expressing *RBP6* at baseline levels, metacyclic-like cells have been generated in the sample (as marked by anti-CRD, yet retained EP procyclin). This could suggest either leaky *RBP6* overexpression, or high amounts of non-specific antibody staining. The leaky overexpression hypothesis may explain why the uninduced samples had drops in cell density consistent with *RBP6* overexpression. Furthermore, the differentiation induced by *RBP6* overexpression does not appear to be fully complete, as all cells are both EP procyclin and CRD positive across the induced and uninduced samples.

19.3 Single cell RNA sequencing data analysis of the *RBP6* overexpression model of *T. brucei* development

19.3.1 Capturing the transcriptome of *RBP6* overexpressing single cells

A staggered *RBP6* induction was carried out to generate three samples on the day of scRNA-seq sample collection: day three, day four and day five post-induced *RBP6* overexpression. The cells from each induced sample were mixed together with a day three:day four:day five cell count ratio of 1:2:2, before scRNA-seq was carried out (Fig.24). More day four and five cells were taken in comparison to day three to ensure a greater chance of capturing metacyclics, as they make up a greater proportion of cells later in the inducement process.

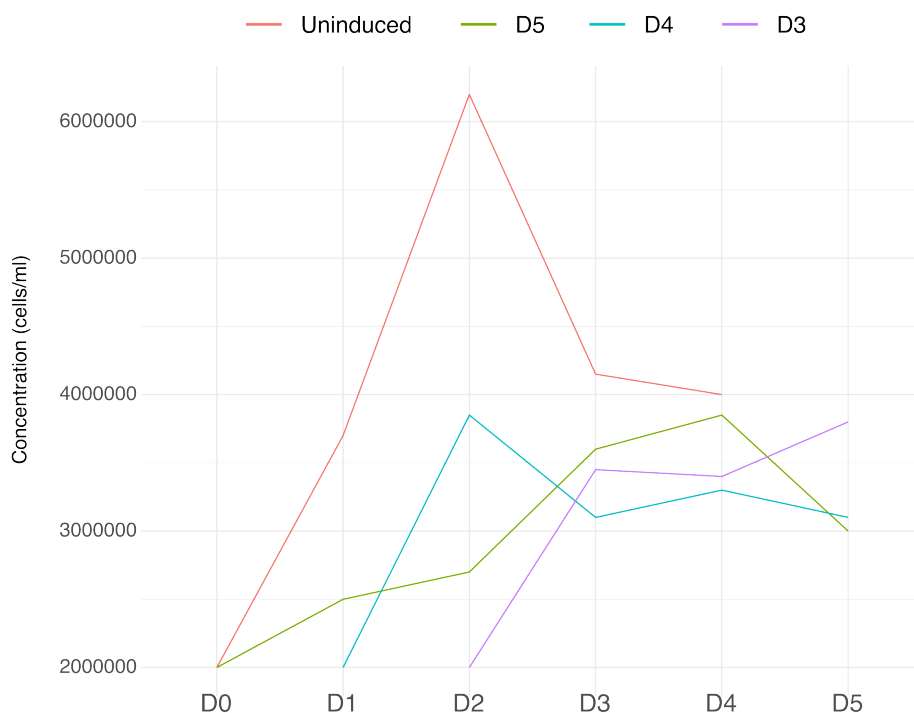


Figure 24: **Growth curves of the *RBP6* overexpressing cells before scRNA-seq**

Concentrations of the *RBP6* overexpressing cells after different lengths of overexpression inducement (three days, four days and five days) as well as an uninduced control. The overexpression-induced cells were collected for scRNA-seq on the fifth day of the experiment. The induced cells were cultured in SDM-80, while the uninduced cells were cultured in SDM-79. Each line represents one culture.

19.3.2 Overview and quality control of the *RBP6* overexpression scRNA-seq dataset

The scRNA-seq data was mapped against a combined reference of the TREU927 genome (with 3' UTRs extended by 2500bp) (Briggs, Rojas, et al. 2021), kDNA genes and mVSG genes (Hutchinson et al. 2021) using the 10X Cellranger Count pipeline. Approximately 16000 cells were sent for sequencing, with the filtered output of 10X Cellranger Count returning 11458 cells, with 1117 median genes returned per cell and 1644 median UMIs per cell. While 90.2% of the reads mapped to the reference genome, only 27.9% of the reads mapped confidently to the genome, signifying most of the reads captured mapped to more than one gene. There was only a drop of 3.8% between the reads confidently mapping to genome and reads confidently mapping to the transcriptome (24.1%). Genes whose reads were captured in less than 10 cells in the dataset were also removed, leaving 8589 genes.

The percentage of reads mapping to genes from the kDNA or ribosomal RNA (rRNA) were then calculated for each cell. The median values of these across the dataset were 1.99% and 1.82%, respectively (Fig.25B, C). These metrics, along with the total number of genes captured per cell, were used to filter out low quality cells. Cells with percentage of kDNA reads more than 5, percentage of rRNA reads more than 4.5, and number of genes per cell less than 400 were removed from the dataset, as these are likely to be stressed, dying or ambient cells (Fig.25B, C). Cells whose number of genes per cell was more than 2500 were also removed, as these could represent doublets. After these quality control steps, 10965 cells

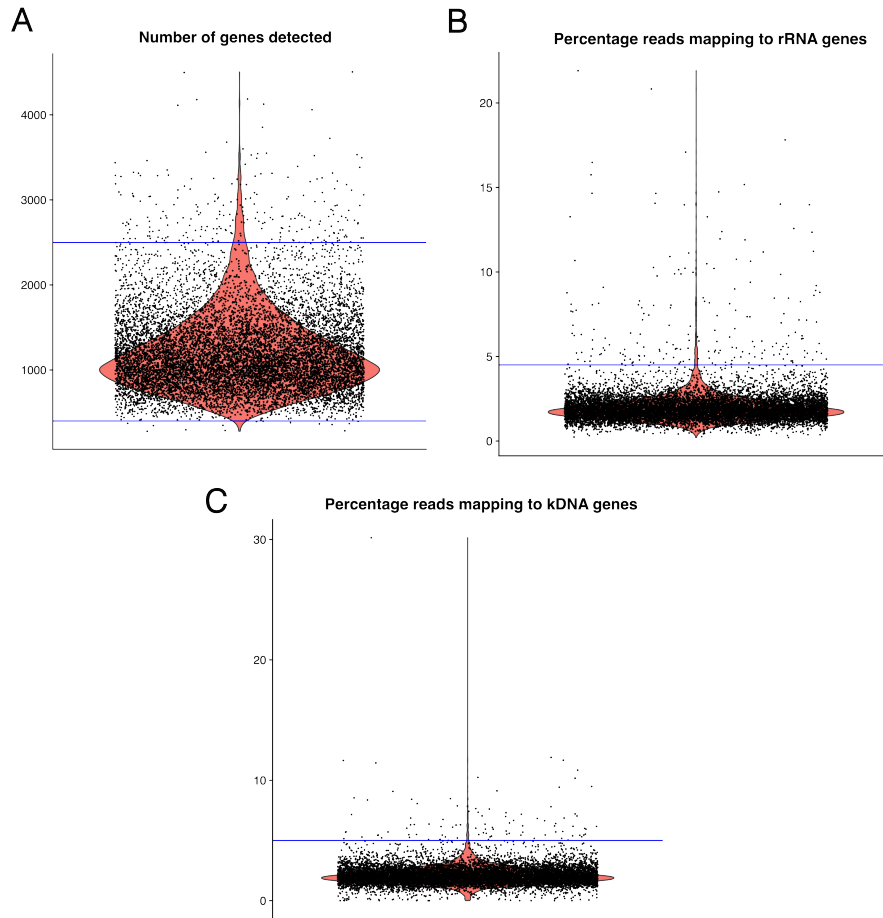


Figure 25: **Quality control metric values for the *RBP6* overexpression scRNA-seq dataset**

Violin plots showing, for each cell in the dataset, the number of genes captured per cell (A), percentage of reads that map to ribosomal RNA genes (B) and percentage of reads that map to kinetoplastid DNA genes (C).

were left in the dataset (Fig.25A).

The count expression values were normalised by the total sum of UMI per cell, $\log(x+1)$ transformed and scaled by the median total counts across all the cells. To choose which genes would make up the feature space of the analysis, the top 2000 highly variable genes were found using Seurat.

The data was then scaled and the effect of total UMI count of a cell regressed out of the scaled expression data. Principal Component Analysis (PCA) was then carried out on the scaled, regressed expression values of the 2000 highly variable expressed genes across all the cells. Louvain clustering was then run on the dataset using the first 12 Principal Components (PCs), with a clustering resolution of 0.6 chosen (generating 9 clusters) (Fig.26A, C). Cells were then embedded in a 2 dimensional space using UMAP, using the first 12 PCs as the basis for UMAP embedding calculation.

Across the insect stage development, cell cycle status changes, with PCFs and epimastigotes being replicative and metacyclics arrested (Taylor et al. 2020). To see what phase of the cell cycle each of the cells in the dataset were in, a method outlined by Briggs and colleagues (Briggs, Marques, et al. 2023) was used to assign a cell cycle phase score to each cell and then assign a phase label depending on the highest score. The cell cycle phases considered were early G1 (G1e), late G1 (G1l), S or G2/mitosis (G2M). The majority of the cells in clusters 2 and 5 were assigned as being in S-phase of the cell cycle, while cluster 4 was dominated by cells which were assigned a G1l label. Cluster 7 had more than 50% of the cells assigned the G1e label and cluster 3 was mainly cells labelled as G2M (Fig.26B, D). The rest of the clusters did not have a particular cell cycle phase identity, making up a majority of the cells (Fig.26B, D).

19.3.3 Regressed analysis of the *RBP6* overexpression scRNA-seq dataset

The above analysis shows that the cell cycle was one of the key determinants of cell clustering and embeddings in the low dimension spaces when analysing the *RBP6* overexpression data (Fig.26B, D). The strong signal of the cell cycle could thus be masking the biological signal of the insect life cycle stage transitions. The scRNA-seq analysis of the *RBP6* overexpression dataset was therefore repeated, with the effect of cell cycle phase score regressed out of the data, and the 130 cell cycle genes that were in the variable feature space removed. PCA was carried out again, with the first 10 PCs being used to generate Louvain clustering (at a resolution of 0.6) and 2D UMAP embeddings (Fig.27A, B, C). The cell cycle label distribution across the clusters was now more uniform, with no cell cycle label being the majority in any cluster. Cluster 5, however, had just under 50% of its cells labelled as G1e phase (Fig.26B, D).

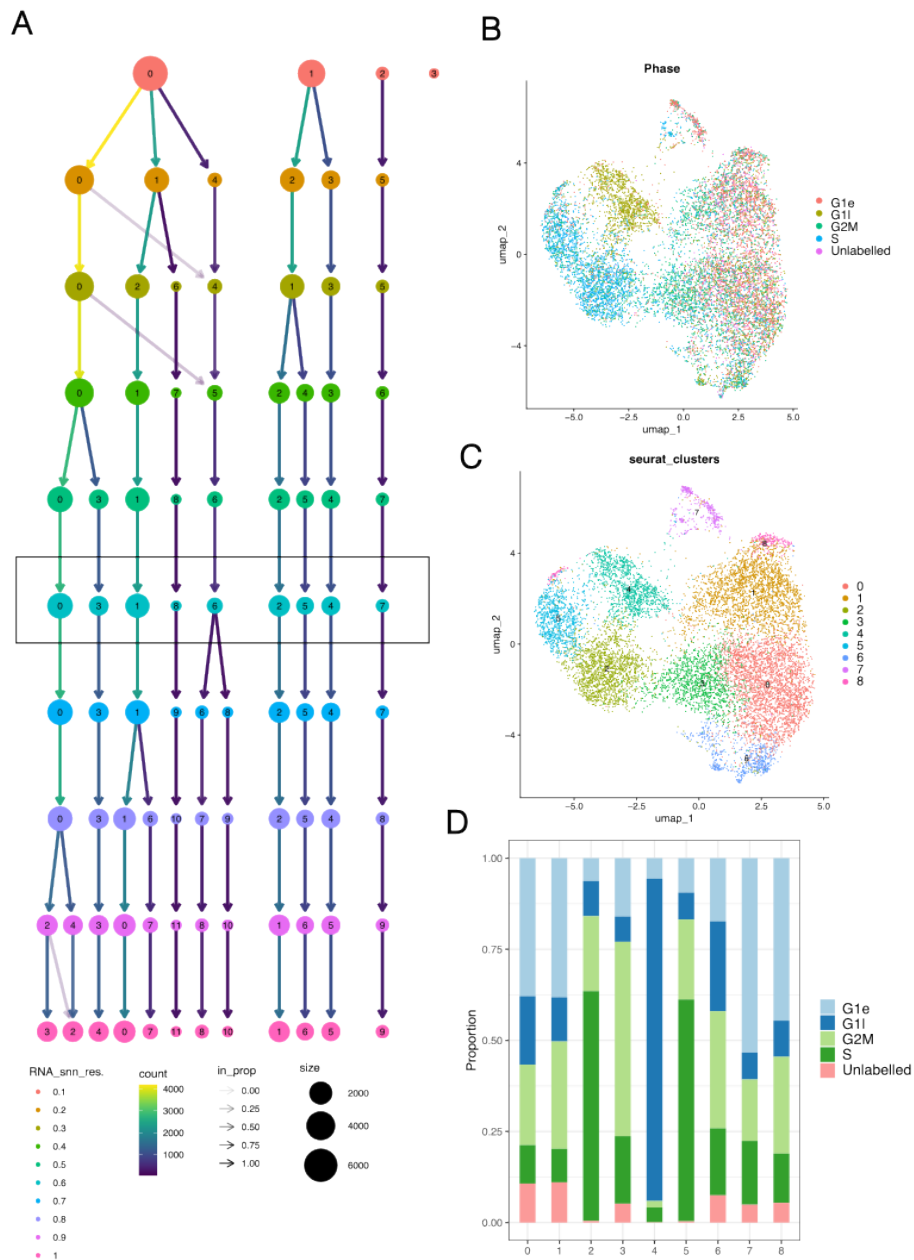


Figure 26: Analysis of the *RBP6* overexpression scRNA-seq dataset with total UMI count regressed out

Clustree plot showing how cells are assigned to clusters across different resolutions for Louvain clustering. The final chosen resolution is outlined in a black rectangle (A). UMAP embeddings of the *RBP6* overexpression scRNA-seq cells coloured by cell cycle phase (B) and cluster identity (C). Proportion plot showing the proportion of cells within each cluster that have a given cell cycle phase annotation (D).

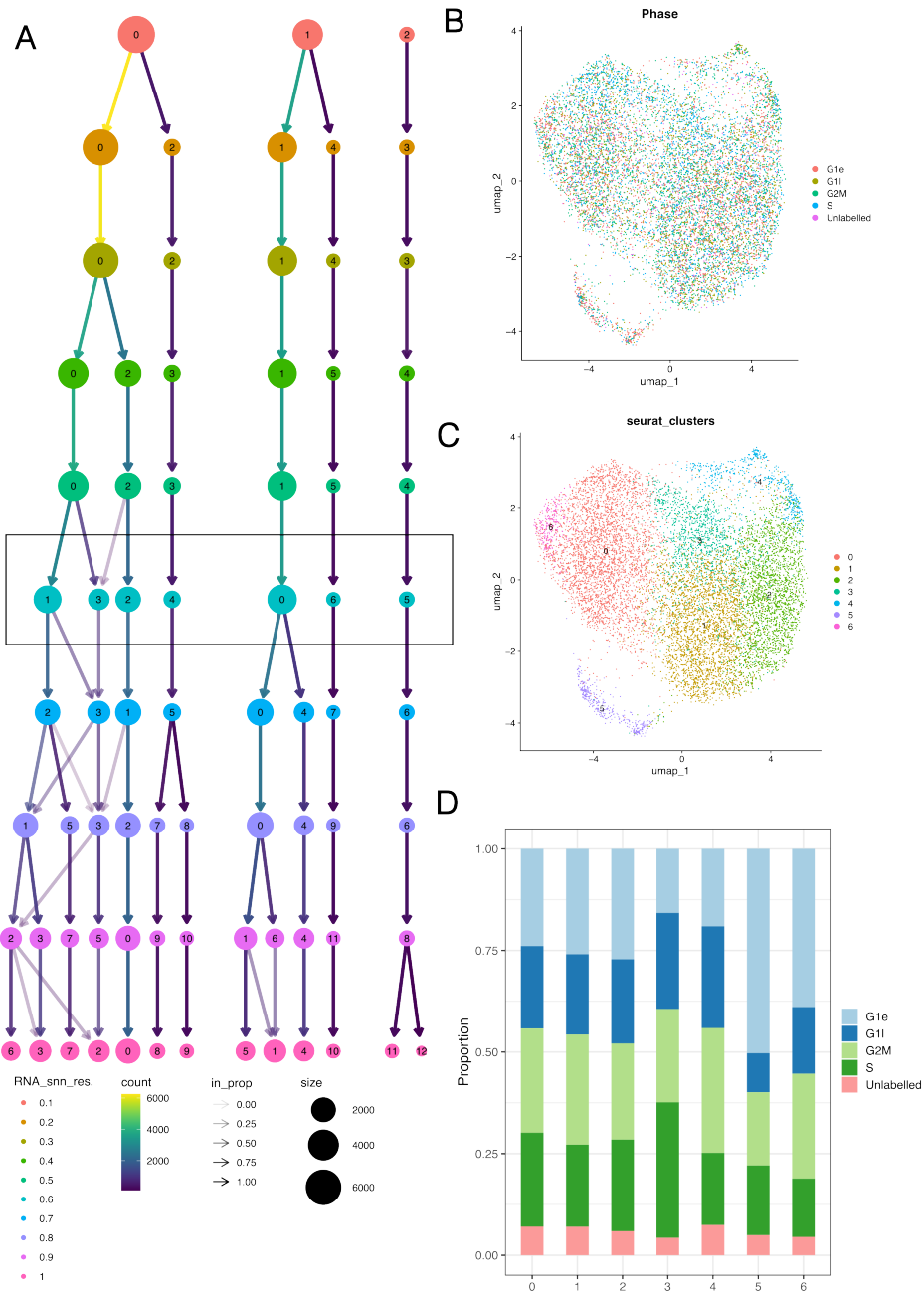


Figure 27: Analysis of the *RBP6* overexpression scRNA-seq dataset with total UMI count and cell cycle phase regressed out.

Clustree plot showing how cells are assigned to a cluster across different resolutions for Louvain clustering. The finalised resolution is outlined in a black rectangle (A). UMAP embeddings of the *RBP6* overexpression scRNA-seq cells coloured by cell cycle phase (B) and cluster identity (C). Proportion plot showing the proportion of cells within each cluster that have a given cell cycle phase annotation (D).

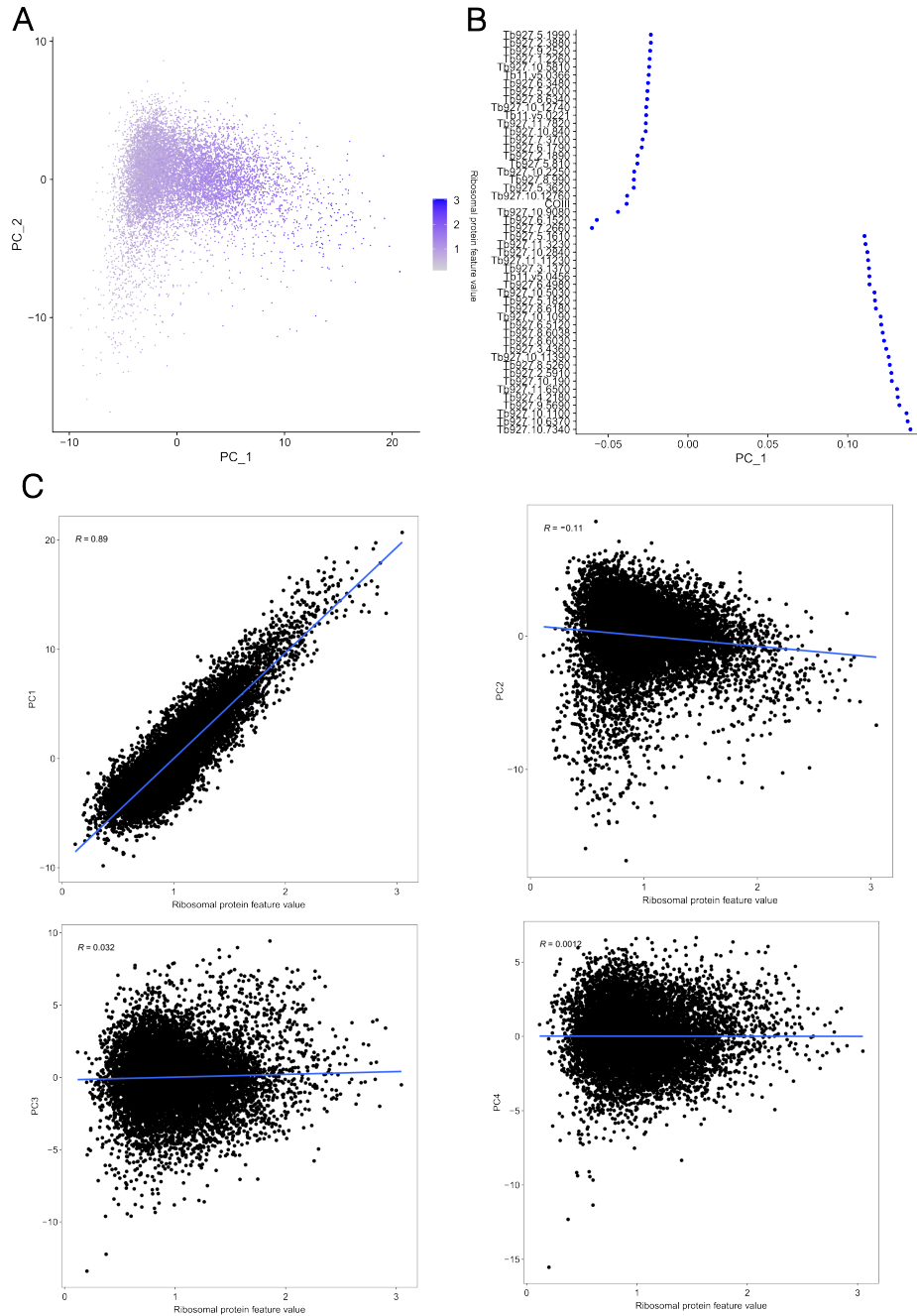


Figure 28: Effect of ribosomal protein encoding gene expression on the PCA embedding of the *RBP6* overexpression scRNA-seq

PCA embeddings of the *RBP6* overexpression scRNA-seq dataset, coloured by the ribosomal protein encoding gene meta feature (A). Dotplot of the top 25 genes most associated with the positive and negative end of PC1 (B). Dotplots showing the relationship between ribosomal protein encoding gene meta feature and PC value (PCs 1-4) for all the cells. Lines represent a linear regression line, with accompanying R value (C).

Cells with high expression of ribosomal protein-encoding genes were located towards the positive side of PC1, while the cells with low expression appeared to be towards the negative side of PC1 (Fig.28A). To see whether ribosomal protein expression explains the positioning of cells across the PCs, the relationship between ribosomal protein-encoding gene meta feature and the first four PCs was assessed. While all the p-values returned were less than 0.05, the R values for PC2, PC3 and PC4 were all close to zero. The relationship between PC1 and ribosomal protein meta feature had an R of 0.89, suggesting a strong correlation between the two variables (Fig.28C). Furthermore, the top 25 genes associated with the positive values of PC1 were all ribosomal protein-encoding genes (Fig.28B). These results suggest that ribosomal protein gene expression is one of the major signals of the *RBP6* overexpression scRNA-seq dataset, and explains the most variation in the cells.

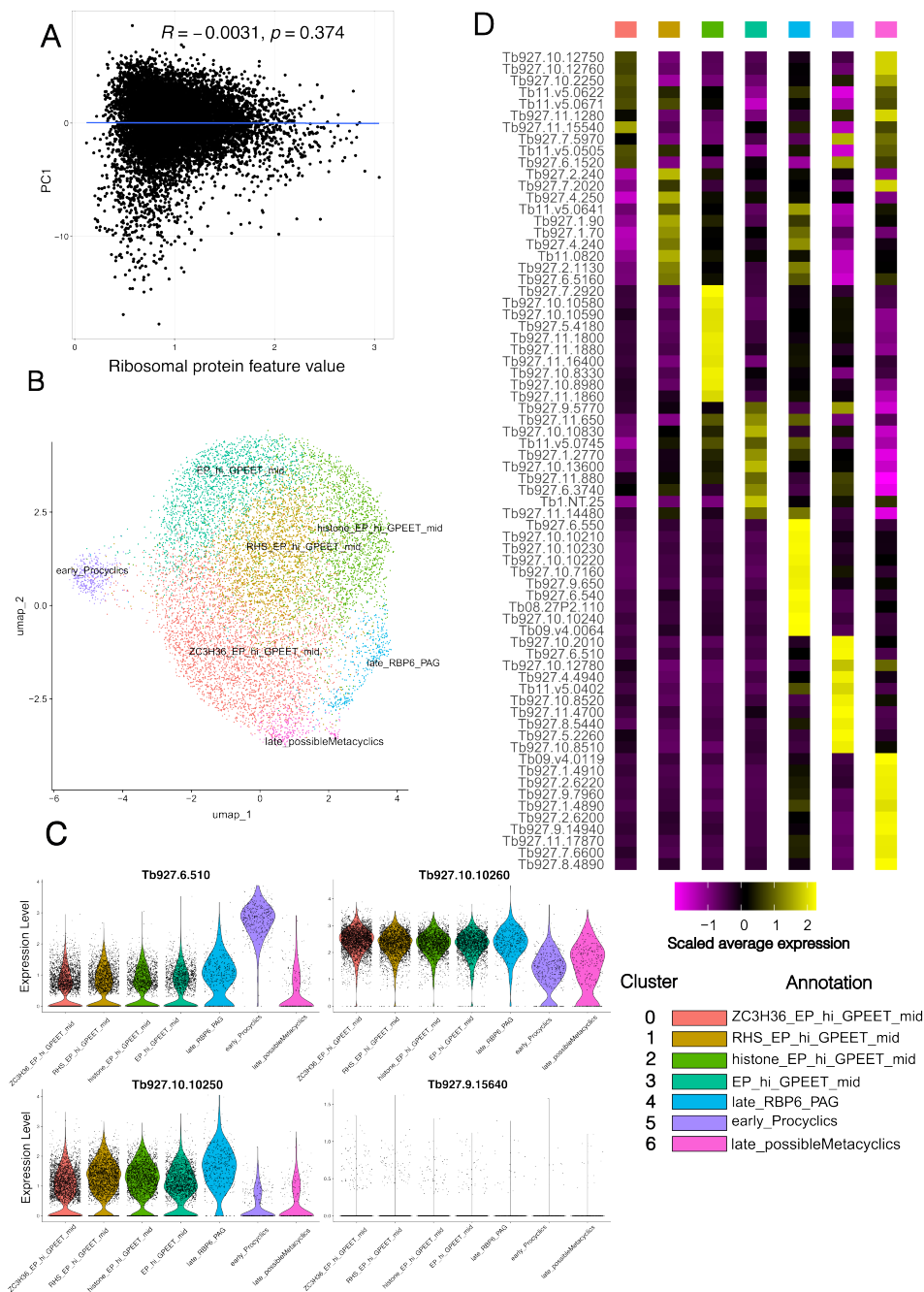


Figure 29: Analysis of the *RBP6* overexpression scRNA-seq dataset with total UMI count, cell cycle phase and ribosomal protein encoding gene expression regressed out.

Dotplot showing the relationship between ribosomal protein encoding gene meta feature and PC 1 for all cells where ribosomal protein gene meta feature scores is regressed out. Line represents a linear regression line, with accompanying R value (A). UMAP embeddings of the *RBP6* overexpression scRNA-seq cells coloured by cluster annotations (B). Violin plots showing the normalised expression of (left to right): GPEET, EP1 procyclin, EP2 procyclin and BARP (C). Heatmap showing the scaled, normalised expression of the top 10 marker genes (in terms of log2FC) averaged across cells in each of the annotated clusters (D). Legend in the bottom right corner shows the cluster number and annotation labels of the clusters as seen in B, C and D.

Given the above data, a new analysis was performed where the effect of ribosomal protein-encoding gene expression, cell cycle and total UMI count was regressed out of the dataset. Regressing out ribosomal protein-encoding gene expression caused a lack of polarisation of cells across PC1, depending on their ribosomal protein encoding gene meta feature (Fig.29A). Seven clusters were generated using this analysis process.

To identify what the different clusters represent in terms of life cycle stage or state, marker gene analysis was performed on the dataset. Marker genes were identified as genes whose absolute \log_2 fold change (\log_2FC) was greater than 0.5, Bonferroni corrected p-value less than 0.05, and where they were expressed in at least 5% of cells in one of the conditions. The markers were then ordered by decreasing \log_2FC . A marker gene of cluster 0 was the zinc finger encoding protein *ZC3H36* (Tb927.10.12760). Nine of the ten top marker genes (in terms of \log_2FC) of cluster 1 encode retrotransposon hot spot (RHS) proteins, specifically five RHS pseudogenes (Tb927.1.90, Tb927.4.240, Tb11.0820, Tb927.2.1130 & Tb927.6.5160), two RHS5 (Tb927.2.240 & Tb927.4.250), an RHS4 (Tb927.1.70) and a RHS7 (Tb927.7.2020), while the other is a gene encoding a putative major vault protein (Tb11.v5.0641) (appendix file 21).

Markers of cluster 2 include many histone encoding genes (Tb927.7.2920, Tb927.10.10580, Tb927.10.10590, Tb927.5.4180, Tb927.11.1800, Tb927.11.1880, Tb927.11.1860), while markers of cluster 3 include a trypanothione peroxidase (Tb927.9.5770) and cyclophilin A (Tb927.11.880) (Fig.29D) (appendix file 21). The main markers of cluster 4 were Procyclin Associated Genes (Tb927.10.10210, Tb927.10.10230, Tb927.10.10220, Tb927.10.7160, Tb927.10.10240) (Fig.29D) (appendix file 21). The cluster 5 markers include hexokinase (Tb927.10.2010), two glucose transporters (Tb927.10.8520, Tb927.10.8510), GPEET procyclin (Tb927.6.510) and *ZC3H37* (Tb927.10.12780) (Fig.29D) (appendix file 21). Finally, cluster 6 markers include VSGs (Tb09.v4.0119, Tb927.11.17870), Expression Site Associated Genes (Tb927.1.4910, Tb927.1.4890), adenosine transporters (Tb927.2.6220, Tb927.2.6200) and *SGM1.7* (Tb927.7.6600) (Fig.29D) (appendix file 21).

ZC3H36 was one of eight transcripts that were found to be upregulated in the mutated *RBP6* overexpression model (compared with uninduced cells), but not in the WT *RBP6* overexpression model (H. Shi, K. Butler, and Tschudi 2018). Given this, the expression values of the other seven transcripts were thus visualised across the *RBP6* overexpression scRNA-seq dataset (Fig.15). None of the other transcripts were more highly expressed in cluster 0 (Fig.15) or were part of its significant marker gene list (appendix file 21). However, Tb927.10.12840 and Tb927.4.4940 (which encode a mitochondrial carrier protein and a hypothetical protein respectively) were more highly expressed in cluster 5 and were significant marker genes of that cluster (Fig.15) (appendix file 21).

SGM1.7 has previously been identified as a transcriptomic marker of *T. brucei* metacyclics (Vigneron et al. 2020) and *GPEET* is a marker of PCFs, which is expressed highly during early stages of tsetse fly infection (Knüsel and Roditi 2013). Other markers of the insect form stages include EP procyclin, which is expressed on epimastigotes and PCFs (Acosta-Serrano et al. 2001), and the BARP family, which are expressed in epimastigotes (Urwylar et al. 2007). When looking at the expression of the most abundantly expressed *BARP* in the dataset (Tb927.9.15640), a distinct lack of localised *BARP* expression was seen across the clusters. The EP procyclins, on the other hand, were expressed highly in clusters 0,1,2,3 and 4, with expression being lower in cluster 6, and the lowest in cluster 5 (Fig.29C). Cluster 5, however, had markedly higher *GPEET* expression than the other clusters (Fig.29C).

There is no publicly available scRNA-seq dataset of the *RBP6* overexpression model of *T. brucei* development, so there are no equivalent datasets to compare these results with. As a proxy for this, the bulk RNA-seq data of the *RBP6* overexpression model generated by Doleželová and colleagues (Dolezelova

et al. 2020) was used to assess the findings of the scRNA-seq data. When looking at the expression of the top 10 (in terms of \log_2FC) marker genes of the scRNA-seq clusters in the bulk RNA-seq data, the markers of clusters 4 and 6 are highest expressed at the day eight timepoint, while the markers of cluster 5 are highest expressed in the uninduced cells (Fig.30A). The marker genes of the other clusters do not have as strong a timepoint expression association, although the expression of cluster 1 markers are lowest expressed in the uninduced cells, and the cluster 2 markers are lowest expressed at the end of the process (Fig.30A). These results validate that the scRNA-seq data contains cells that have a similar transcriptomic profile at the start and end points of the *RBP6* overexpression process. While the expression of EP2 procyclin and *SGM1.7* increase in expression over the *RBP6* induction captured by the bulk RNA-seq data, BARP expression in the induced samples appears to be static, only increasing from the levels seen in the uninduced samples (Fig.30B, C & D).

To assign labels to the clusters, the following annotation strategy was carried out, where clusters were annotated based on the marker genes of the clusters and the expression of *GPEET*, EP procyclins 1 and 2, as these represent defined markers of the insect stage life cycle. Clusters that had medium expression (relative to the other clusters) of *GPEET* and high expression of EP procyclins were annotated as such, as well as by the results of their marker gene analysis. For example, cluster 1 was annotated as "RHS_EP_hi_GPEET_mid" as many of its marker genes encoded RHS proteins, and it had high expression of EP procyclins and medium expression of *GPEET* (Fig.29B, D). Clusters 4 and 6 were defined as "late" as their marker genes were highly expressed in the late stages of the *RBP6* overexpression bulk RNA-seq datasets, with cluster 4 being marked additionally by its high expression of PAG genes (Fig30A, Fig.29B, D). As cluster 5 was marked by the high expression of *GPEET* and low expression of EP procyclins, so they were annotated as "early_Procyclics" (Knüsel and Roditi 2013) (Fig.29B, C). The high *GPEET* expression and the marker genes were the highest expressed in the uninduced bulk RNA-seq data (Fig.30A), suggesting that cluster 5 represents PCF cells that are unresponsive to overexpression stimuli, or are unable to overexpress the protein.

19.4 Integration of *RBP6* overexpression cells and *in vivo* insect life cycle stage cells

Given the that the marker genes of cluster 6 were highest expressed at the end of the bulk RNA-seq timecourse and included VSG transcripts and *SGM1.7*, the likely conclusion drawn would be that these cells contain metacyclic forms. When the individual expression values for each cell in the dataset are plotted, however, these genes were expressed at similar levels in most of the *RBP6* overexpression dataset clusters, with a few outliers. For example, expression of *SGM.17* was at a similar level to the other clusters, except for one cell, which had high expression of the gene (Fig.31A). Furthermore, when the expression of the mVSGs were plotted, some cells with high expression of these genes could be seen in cluster 6 (Fig.31B). Together, these results suggest that very few metacyclics have been captured in the scRNA-seq data.

To try to confirm whether or not metacyclics were present in the scRNA-seq dataset, the *RBP6* overexpression scRNA-seq dataset was integrated with a scRNA-seq dataset of *in vivo* insect life cycle stages. There are three published scRNA-seq datasets that capture *T. brucei* cells directly from the insect. Howick and colleagues (Howick, L. Peacock, et al. 2022) captured cells from the tsetse fly's proventriculus, midgut and salivary glands using a modified protocol of the plate based SMART-seq2 method (Picelli et al. 2014). The SMART-seq2 data contains full length transcript, non-UMI reads, while the *RBP6* overexpression dataset contains 3' enriched UMI reads. As different scRNA-seq capture technologies can introduce large batch effects (Luecken and Theis 2019), the *RBP6* overexpression dataset

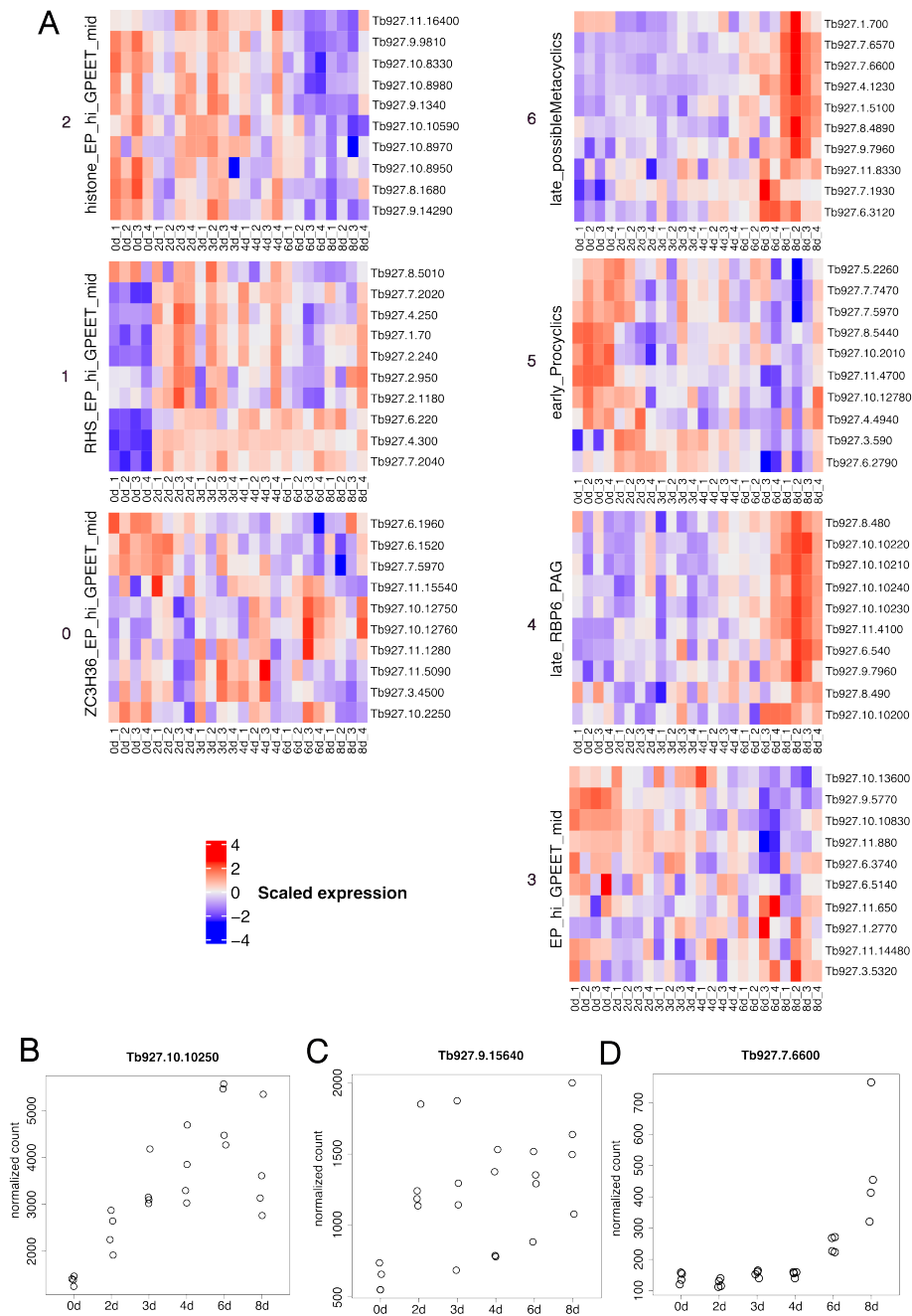


Figure 30: Comparison of scRNA-seq and bulk RNA-seq datasets of *RBP6* overexpressing *T. brucei* cells

Heatmap of the scaled, normalised gene expression of the top 10 marker genes (in terms of log₂FC) of the annotated clusters as seen in Fig.29 across the bulk RNA-seq datasets from Doleželová and colleagues (Dolezelova et al. 2020) that contain *RBP6* overexpressing *T. brucei* cells across two (2d), four (4d), six (6d) and eight (8d) days of *RBP6* overexpression, as well as uninduced cells (0d). The annotated cluster names as well as cluster ID for the marker lists are given to the left of the heatmaps.

(A). Normalised expression of EP 2 Procyclin (B), BARP (C) and SGM1.7 (D) across the bulk RNA-seq datasets of *RBP6* overexpressing and uninduced *T. brucei* cells.

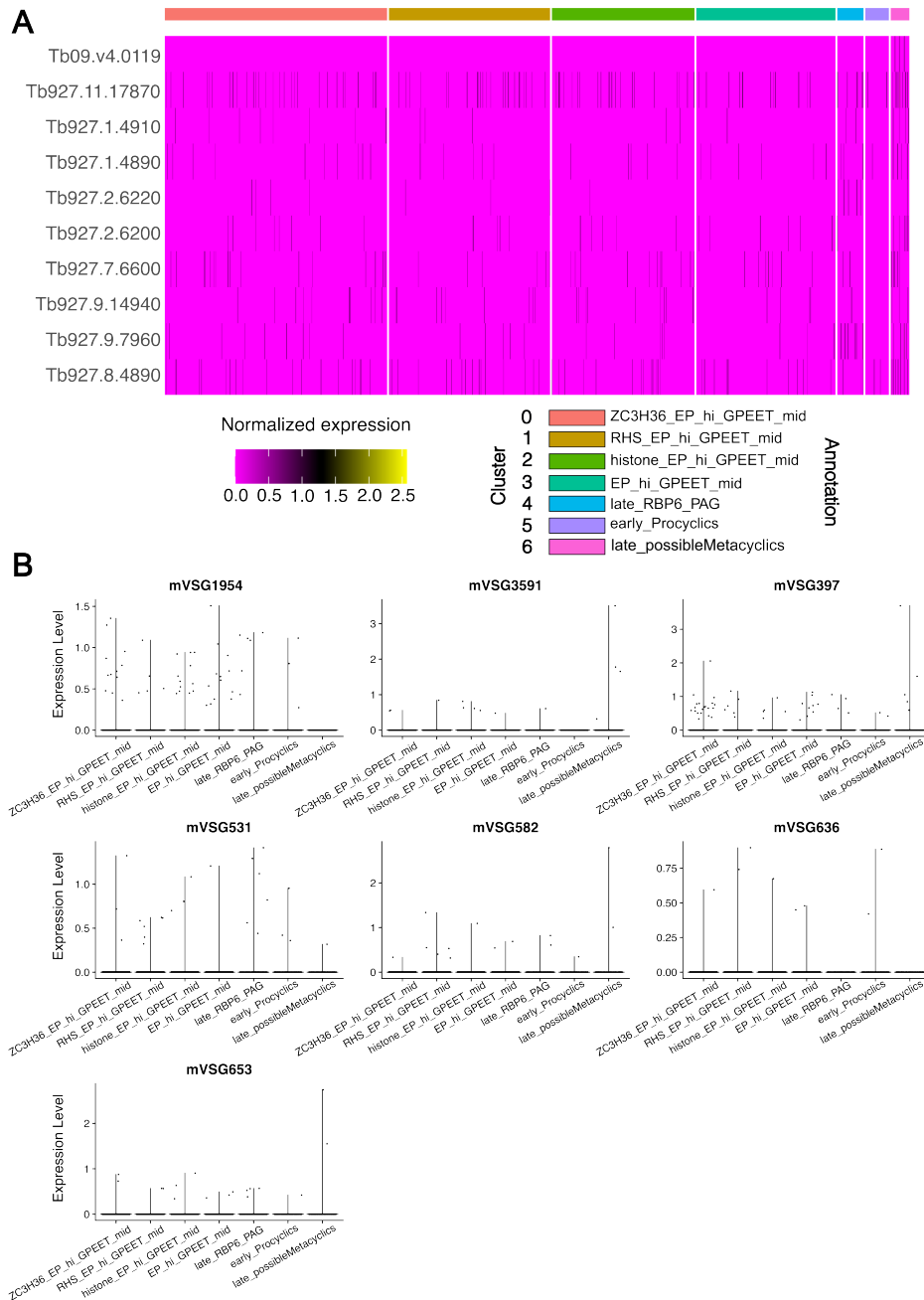


Figure 31: Expression values of cluster 6 marker genes plotted at the single cell level

Heatmap showing the normalised expression values for each cell of the top 10 (in terms of log₂FC) significant marker genes of cluster 6 (A). Violin plots showing the normalised gene expression of the mVSG genes (B).

was not integrated with this dataset. Vigneron and colleagues (Vigneron et al. 2020) have data with 3' enriched UMI reads, but they only captured cells in the salivary glands of the tsetse fly, and they did not identify any midgut PCFs. Finally, while Hutchinson and colleagues (Hutchinson et al. 2021) only sampled parasites from the salivary gland (identifying mVSG expressing cells), they detected midgut stage parasites in their data, and they also used a 3' enriched UMI reads based sequencing method. Thus, the *RBP6* overexpression scRNA-seq and Hutchinson scRNA-seq datasets were integrated using Harmony (Korsunsky et al. 2019) integration (Fig.32A & C, Fig.33A & B).

The clusters split up along the lines of the dataset origin, showing little mixing of the datasets. Furthermore, some of the resulting cluster groupings did not reflect the underlying biology. For example, the cells from the *in vivo* insect stage life cycle dataset marked "Midgut forms", "Metacyclics" and "Pre-metacyclic" grouped together in the integrated cluster 5, despite the midgut forms and metacyclics being at the opposite points of the insect stage life cycle development (Fig.18, Fig.32D, Fig.33A & B). From the Clustree it can be observed that clusters mainly containing *in vivo* insect stage life cycle cells and the clusters that mainly contain *RBP6* overexpression cells did not share a common lineage, suggesting a lack of transcriptomic similarity between the cells (Fig.32C). As expected, cluster specific *BARP* expression was seen in the *in vivo* insect stage life cycle dataset epimastigote clusters (Fig.32B).

While the possible metacyclic cluster cells in the *RBP6* overexpression dataset did not share a lot of overlap with the "Pre-metacyclic" and "Metacyclics" clusters from the *in vivo* insect stage life cycle dataset, the overlap of the "Midgut forms" and "Pre-metacyclic" and "Metacyclics" call into question the accuracy of the integration. Thus, two more integration methods were attempted, Seurat V5 CCA (Y. Hao, T. Stuart, et al. 2024) and fastMNN (Haghverdi et al. 2018). For the Seurat V5 CCA integration, the *RBP6* overexpression metacyclics and the *in vivo* metacyclics share a common lineage in the Clustree (Fig.34C), but they separate from each other early in the tree. The "early_Procyclics" from the *RBP6* overexpression mostly form their own independent lineage, with little crossover with any other life cycle stage (Fig.34D). In the fastMNN integration, the possible metacyclic cells from the *RBP6* overexpression and *in vivo* metacyclics were contained within cluster 4, and there was some overlap between the "Midgut forms" and the "early_procyclics" (Fig.34). While there is an expectation that the "Midgut forms" and the "early_procyclics" cells would overlap, there is a possibility that the "Midgut forms" cells represent later PCFs as they have higher expression of EP procyclins than the "early_procyclics" cells and have lower *GPEET* expression (Knüsel and Roditi 2013) (Fig.32B). The epimastigotes from the *in vivo* insect life cycle stage dataset mostly cluster on their own (Fig.34B). This is consistent with the lack of *BARP* expression in the *RBP6* overexpression cells (Fig.32B), which may confirm a lack of epimastigote transcriptomic signature cells in the *RBP6* overexpression dataset.

The lack of consistent integration of the *in vivo* insect life cycle stage dataset and the *RBP6* overexpression dataset makes it difficult to assess whether or not the possible metacyclics cluster in the *RBP6* overexpression dataset contains metacyclics. To investigate the differences between the *in vivo* metacyclics and the *in vitro* cells in the "late_possibleMetacyclics" cluster, the DE genes (absolute $\log_2FC > 0.5$, Bonferroni corrected p-value < 0.05 & expressed in at least 10% of the cells in one of the cluster conditions) were found between the two. 2167 genes were identified as being significantly DE between the possible metacyclic *RBP6* overexpression clusters and the metacyclics cluster from the *in vivo* insect life cycle stage dataset, with 1891 genes significantly higher expressed in the *RBP6* overexpression dataset cells and 276 genes significantly higher expressed in the *in vivo* insect life cycle stage dataset cells (appendix file 22).

DE genes significantly higher expressed in the *in vivo* insect life cycle stage dataset metacyclics include several invariant surface glycoproteins (ISG) (*ISG64*:Tb927.5.1410, *ISG65*:Tb927.2.3270 & *ISG75*:Tb927.5.350),

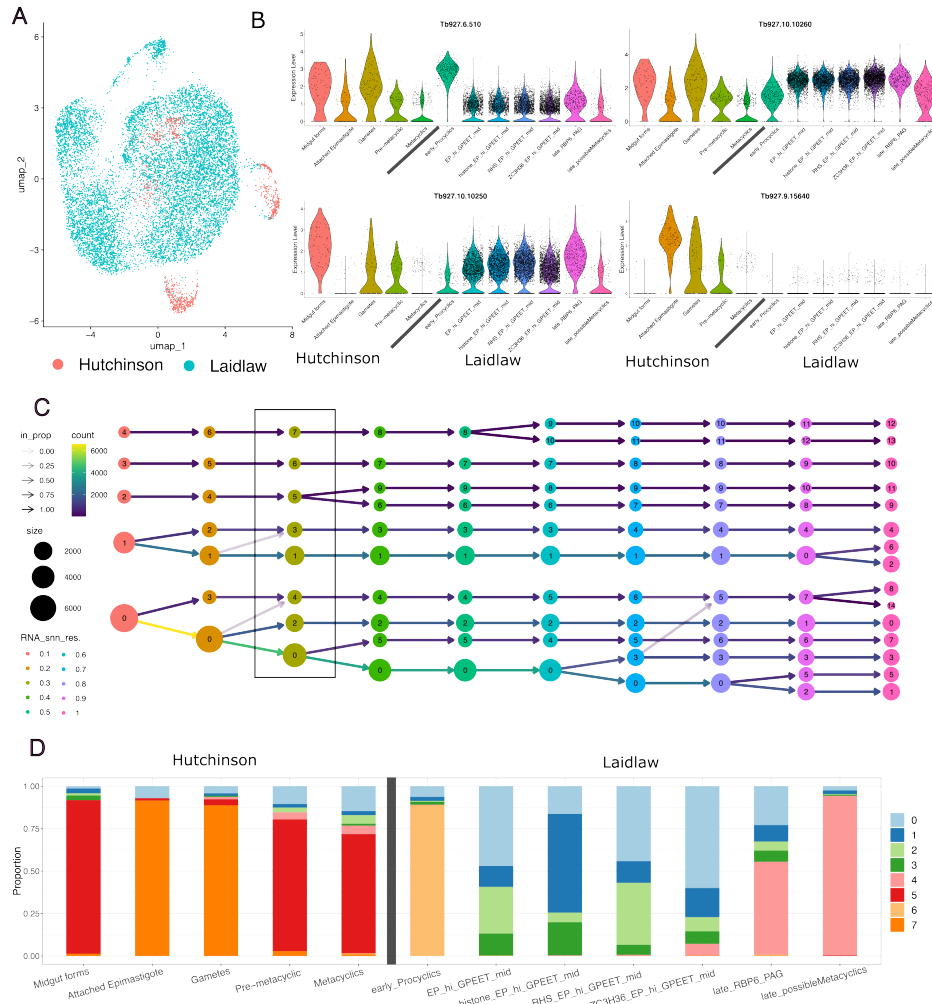


Figure 32: Results of the harmony integration of the *in vivo* insect stage cells and the *RBP6* overexpression generated cells.

UMAP embedding of the harmony integrated datasets, with the cells coloured by the dataset they come from (A). Violin plots showing the normalised expression of GPEET procyclin (Top left), EP1 procyclin (Top right), EP2 procyclin (bottom left) and BARP (bottom right) (B). Clustree plot showing the cell changes between clusters across the different resolution settings. The chosen resolution for the harmony integration is surrounded in a black box (C). Proportion plot showing the proportional distribution of the cell types identified in the *RBP6* overexpression and *in vivo* insect stage cell datasets across the clusters of the integrated dataset (D). The black lines on B and D are used to separate the original clusters by whether they are from the Hutchinson (*in vivo* insect life cycle stage cells) or Laidlaw (*RBP6* overexpression cells) datasets.

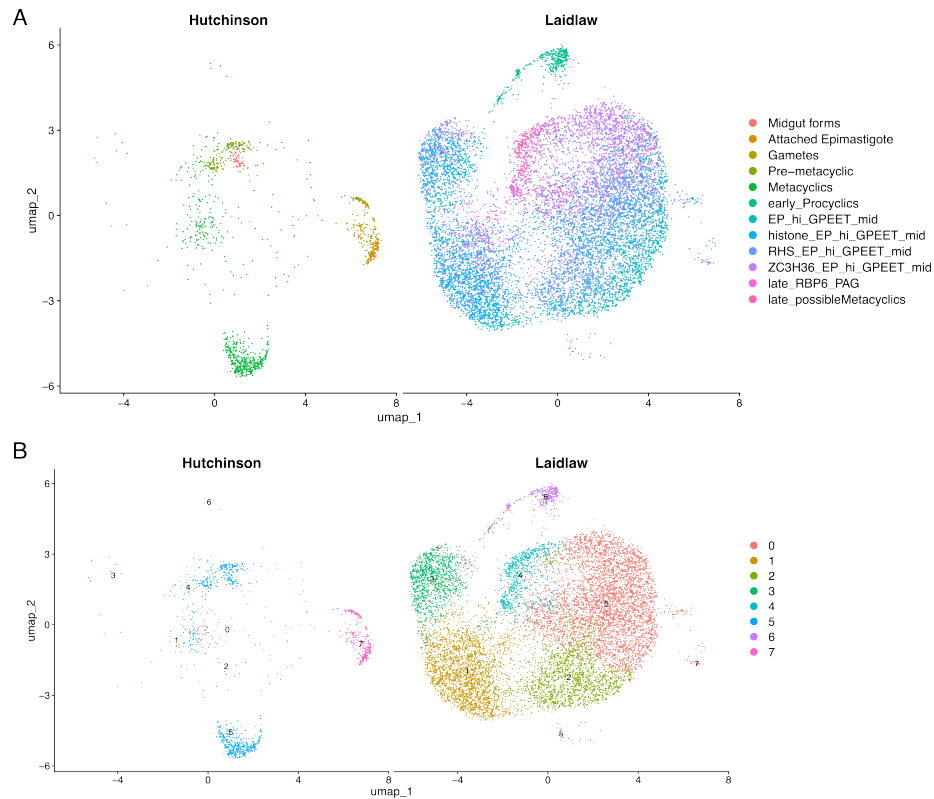


Figure 33: **Split UMAP** results of the harmony integration of the *in vivo* insect stage cells and the *RBP6* overexpression generated cells.

UMAP embedding of the harmony integration, split by whether the cells are from the Hutchinson (*in vivo* insect life cycle stage cells) or Laidlaw (*RBP6* overexpression cells) datasets. Cells are coloured by original cell type IDs (A) and new integrated clustering IDs (B).

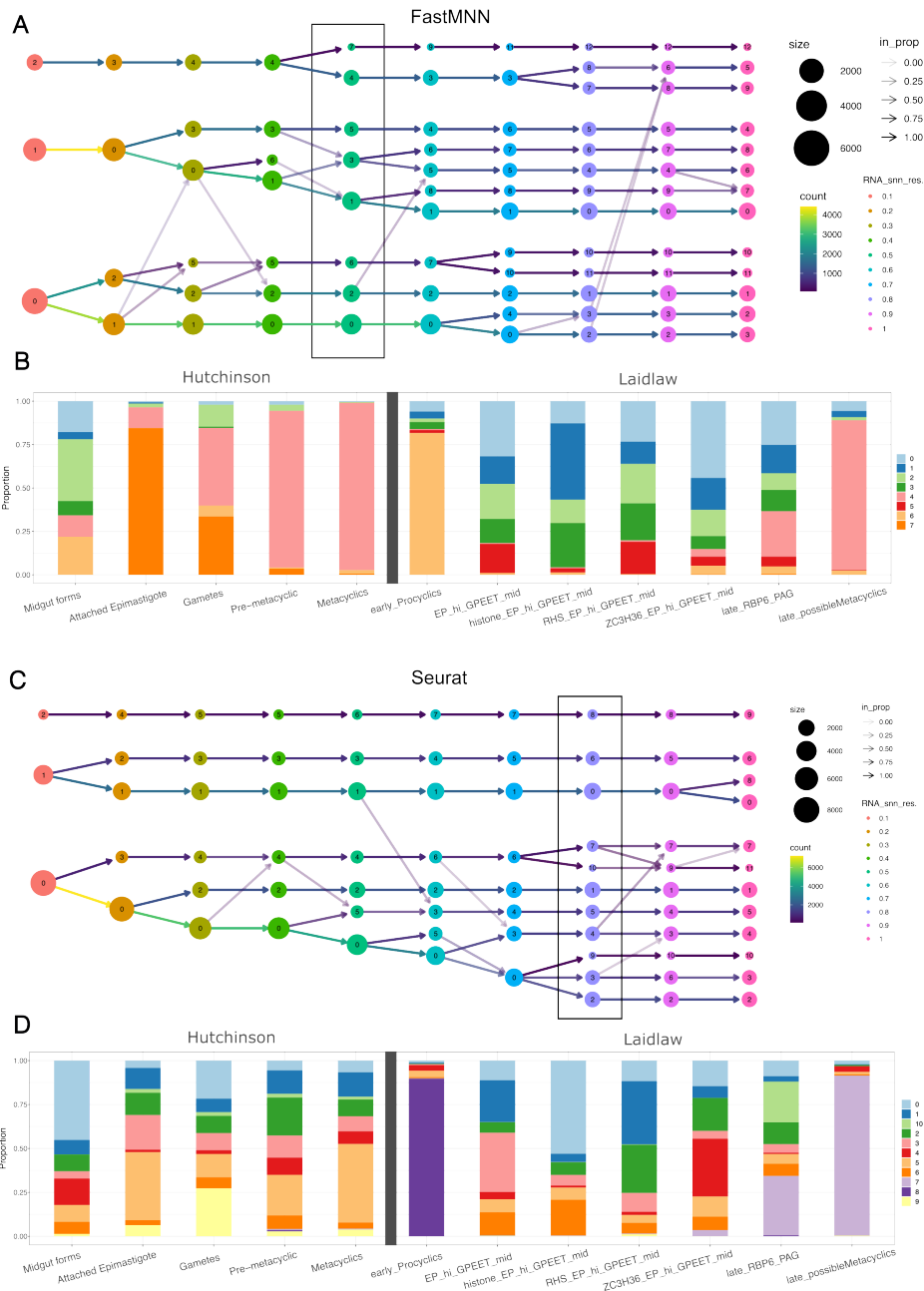


Figure 34: Integration results of the *in vivo* and *RBP6* overexpression *T. brucei* datasets using fastMNN and Seurat

Clustree plot showing the cell changes between clusters across the different resolution settings for the fastMNN (A) and Seurat (B) integrations. The chosen resolution for the integrations is surrounded in a black box. Proportion plot showing the proportional distribution of the cell types identified in the *RBP6* overexpression and *in vivo* insect stage cell datasets across the clusters of the fastMNN (B) and Seurat (D) integrated object. The black lines on B and D are used to separate the original clusters by whether they are from the Hutchinson (*in vivo* cells) or Laidlaw (*RBP6* overexpression cells) datasets.

a glycosomal phosphofructokinase (Tb927.3.3270), *SGM1.7* (Tb927.7.6600), alternative oxidase (Tb927.10.7090) and a succinate dehydrogenase (Tb927.9.5960) (appendix file 22). As many of these genes seem to play a role in metabolism, metabolic pathway enrichment was carried out using TriTrypDB with KEGG and MetaCyc pathways (Caspi et al. 2014 & Kanehisa et al. 2017), with many of the significantly enriched (Benjamini-Hochberg adjusted p-value < 0.05) pathways associated with glycolysis found (appendix file 23).

DE genes significantly higher expressed in the "late_possibleMetacyclics" cluster cells include *PDH* (Tb927.7.210) and *DP5CDH* (Tb927.10.3210), glutamate dehydrogenase (Tb927.9.5900), *LDH* (Tb927.6.2790), repressor of differentiation kinase 2 (Tb927.4.5310) and cytochrome oxidase subunits (Tb927.1.4100, Tb927.9.3170, Tb927.10.280 & Tb927.3.1410). Metabolic pathway enrichment analysis revealed significant enrichment of TCA cycle-associated pathway genes in the DE gene list (appendix file 24).

Christiano and colleagues (Christiano et al. 2017) used the *RBP6* overexpression system to generate metacyclics and identify transcriptomic differences between metacyclics and PCFs. All the DE genes mentioned in the previous paragraph with significantly higher expression in the *in vivo* insect life cycle stage dataset cells were found by the authors to be upregulated in metacyclics versus the PCFs at the transcript level. Furthermore, many of the genes (specifically *PDH*, *DP5CDH*, glutamate dehydrogenase and cytochrome oxidase subunits) that were significantly higher expressed in the possible metacyclics from the *RBP6* overexpression dataset, were identified by the authors as being higher expressed in PCFs compared to metacyclics, at the transcript level.

Metacyclics are said to be preadapted to the host environment, and one of the ways in which this occurs is by changing their metabolism from mainly oxidative phosphorylation to glycolysis-based energy system (Christiano et al. 2017). The enrichment of glycolysis metabolic pathway genes in the *in vivo* metacyclics specific DE genes reflect this preadaptation, while the enrichment of TCA genes in the *RBP6* overexpression possible metacyclics specific DE gene list, show that this change in metabolism is not present, at least not to the same degree. These results indicate that the possible metacyclic cluster in the *RBP6* overexpression dataset are not transcriptionally similar to *in vivo* metacyclics and thus may not represent metacyclics.

19.5 Integration of *RBP6* overexpression cells and *in vitro* culture PCFs

These results lead to an alternative hypothesis that the overexpression failed, and these cells instead represent PCFs. To test this hypothesis, the *RBP6* overexpression dataset was integrated with a scRNA-seq of fresh *in vitro* SDM-79 cultured PCFs (Briggs, Marques, et al. 2023) (Fig.35A, B, Fig.36A, B). The results of the Harmony integration shows that less than 5% of the *in vitro* PCFs co-cluster with either the "early_Procyclics" or the possible metacyclics cluster (Fig.35D). While the expression values of EP1 and EP2 procyclin are comparable between the *in vitro* PCFs and the "late_possibleMetacyclics" cluster, *GPEET* expression is higher in the *in vitro* PCF cells; with expression levels comparable to that of the "early_Procyclics" cluster cells (Fig.35C, Fig.36A, B).

When integration was repeated with Seurat V5 CCA and fastMNN, a lack of consistent integration was seen, similar to the integration of the *RBP6* overexpression and *in vivo* insect life cycle stages dataset (Fig.37). The Seurat integration had 93.6% of "early_Procyclics" cells forming a unique cluster (Fig.37C, D), while the fastMNN integration had 46.7% of these cells mixed in with other cells in the *RBP6* overexpression dataset (Fig.37A, B). While the Seurat integration had the possible metacyclics and the *in vitro* PCFs sharing a common clustering ancestor (Fig.37C), both Harmony and fastMNN

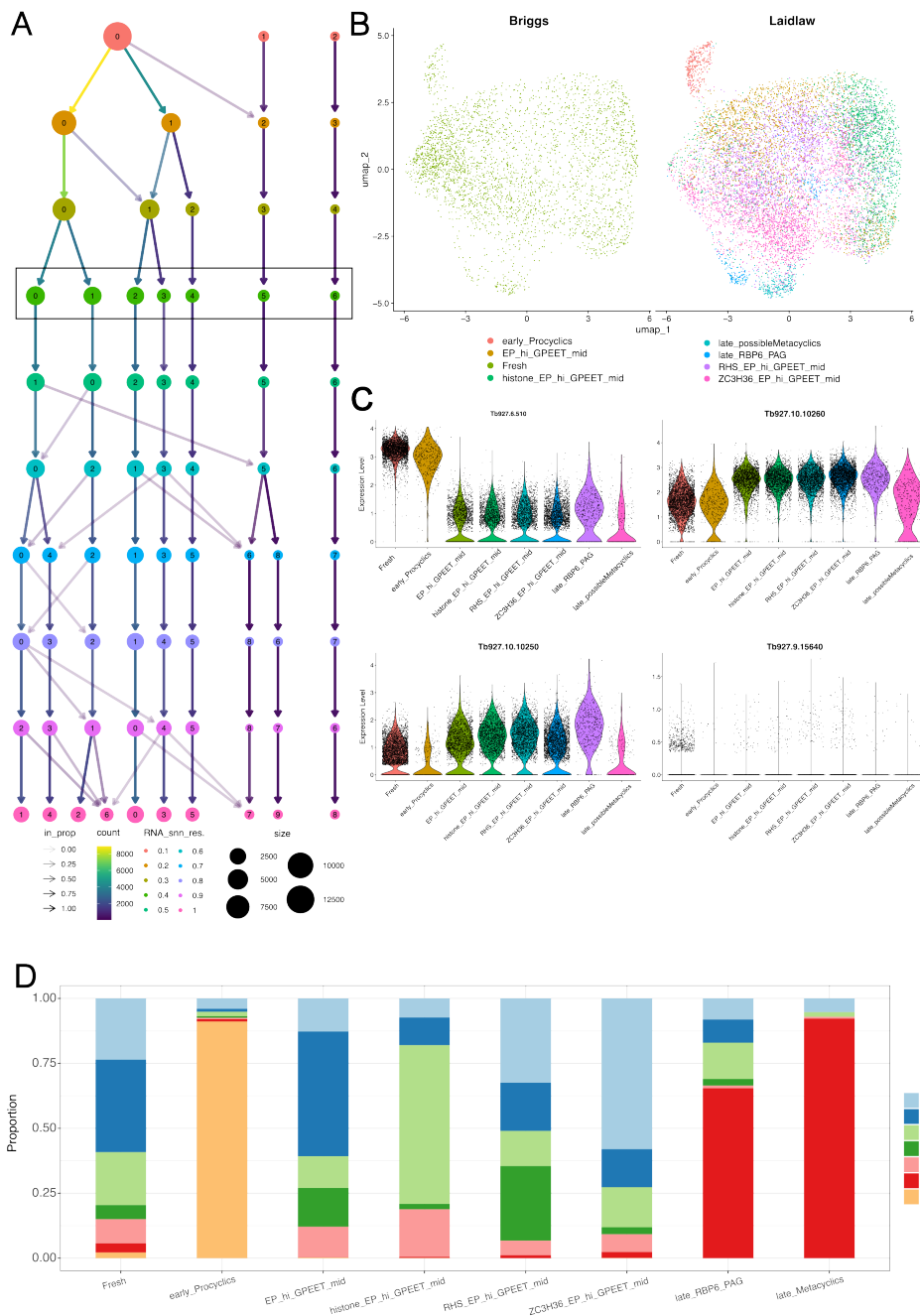


Figure 35: Integration and comparison of the *in vitro* PCFs and *RBP6* overexpression *T. brucei* datasets

UMAP embedding of the harmony integrated datasets, with the cells coloured by the dataset they come from (A). Violin plots showing the normalised expression of *GPEET* (Top left), EP1 procyclin (Top right), EP2 procyclin (bottom left) and *BARP* (bottom right) (B). Clustree plot showing the cell changes between clusters across the different resolution settings. The chosen resolution for the harmony integration is surrounded in a black box (C). Proportion plot showing the proportional distribution of the cell types identified in the *RBP6* overexpression and *in vitro* PCF datasets across the clusters of the integrated dataset (D). The black lines on B and D are used to separate the original clusters by whether they are from the Briggs (*in vitro* PCFs) or Laidlaw (*RBP6* overexpression cells) datasets.

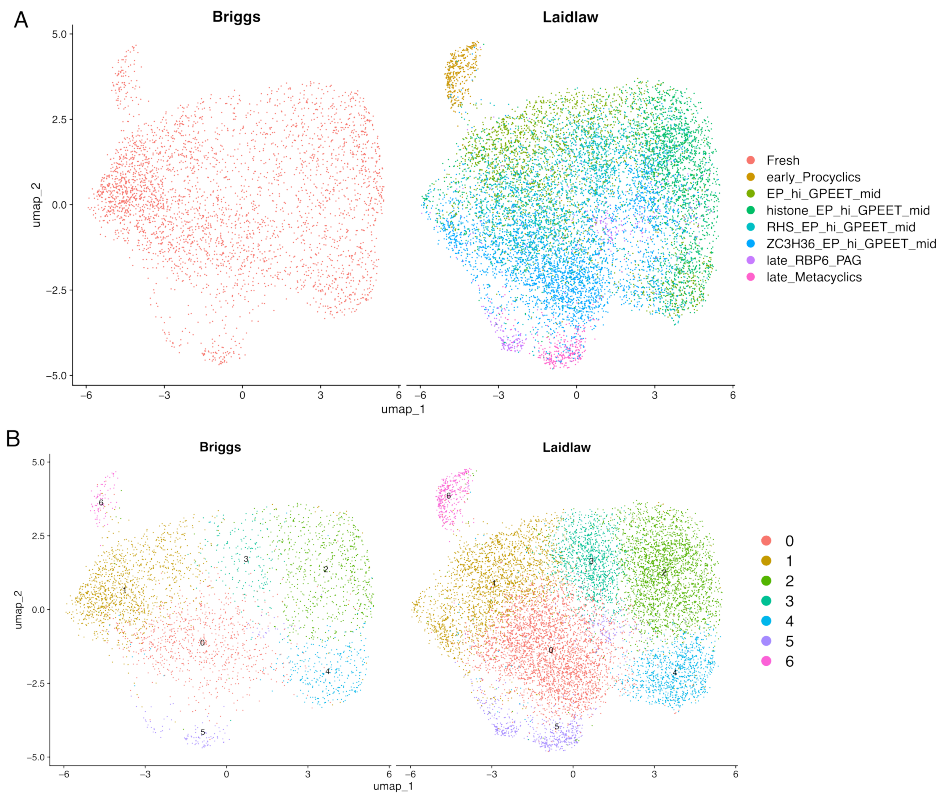


Figure 36: **Split UMAP results of the harmony integration of the *in vitro* PCFs and *RBP6* overexpression *T. brucei* datasets**

UMAP embedding of the harmony integration, split by whether the cells are from the Briggs (*in vitro* PCFs) or Laidlaw (*RBP6* overexpression cells) datasets. Cells are coloured by original cell type IDs (A) and new integrated clustering IDs (B).

had most of these cells being unrelated to the *in vitro* PCFs (Fig.35A, Fig.37A).

As the results of the integration showed disagreement over the relationship between the *RBP6* overexpression dataset possible metacyclics and the *in vitro* derived PCFs, the DE genes between these clusters was identified in a similar fashion to the analysis carried out for the *in vivo* metacyclics versus the *RBP6* overexpression possible metacyclics.

In total, 1946 significant DE genes were identified, with 1703 DE genes found to be significantly more highly expressed in the *in vitro* PCFs and 243 DE genes found to be significantly higher expressed in the *RBP6* overexpression possible metacyclics (appendix file 25). These genes include *PDH* (Tb927.7.210), *DP5CDH* (Tb927.10.3210), glutamate dehydrogenase (Tb927.9.5900) and *LDH* (Tb927.6.2790), all of which are higher expressed in PCFs compared to *RBP6* overexpression metacyclics (Christiano et al. 2017) (appendix file 25). No metabolic pathway genes were significantly enriched in either the *in vitro* PCFs or *RBP6* overexpression "late_possibleMetacyclics" cells DE gene lists (appendix file 26 & 27). These results suggest that the "late_possibleMetacyclics" cells are not transcriptionally similar to *in vitro* PCFs, and thus may represent *RBP6* overexpression generated metacyclics.

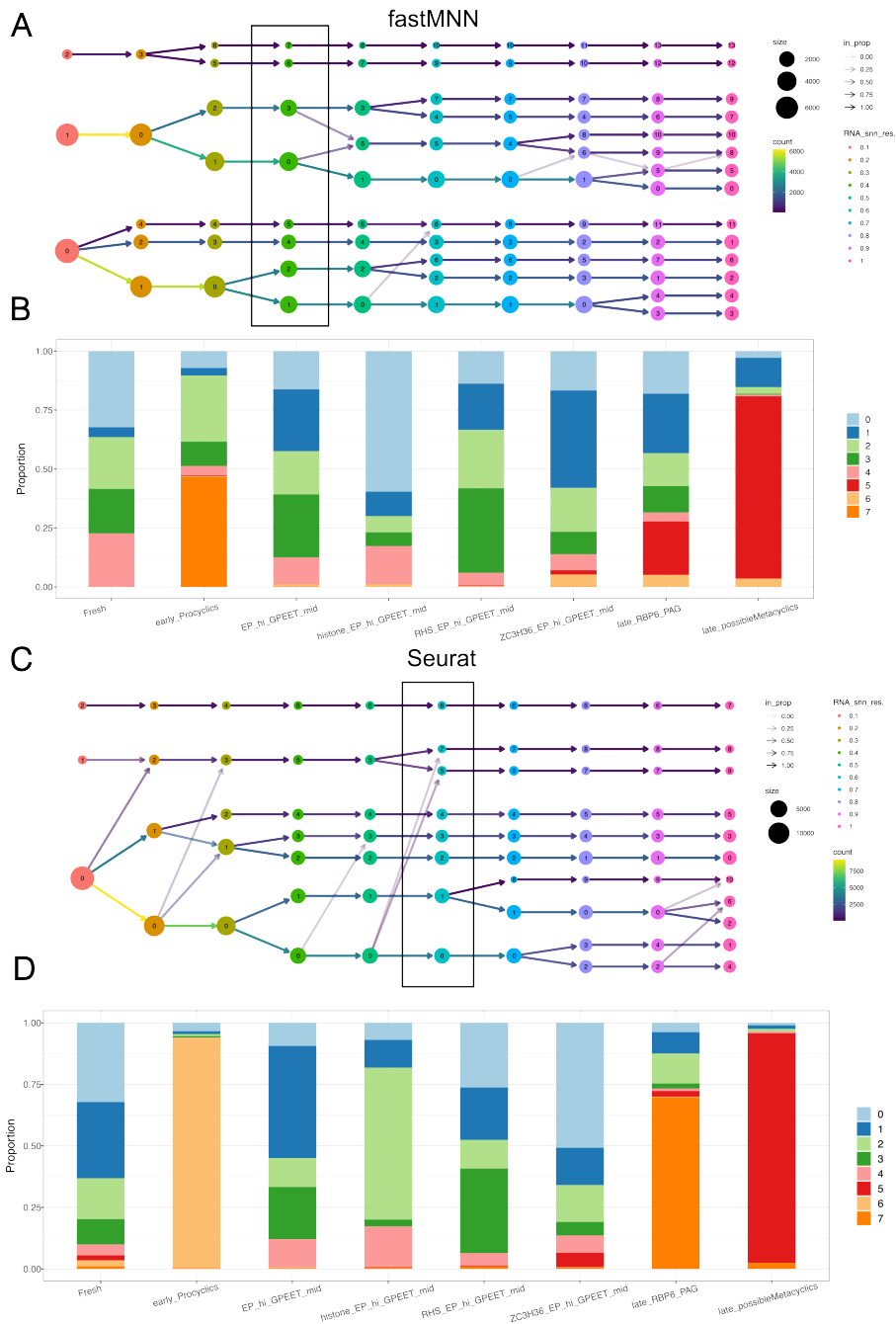


Figure 37: Integration results of the *in vitro* PCFs and *RBP6* overexpression *T. brucei* datasets using fastMNN and Seurat

Clustree plot showing the cell changes between clusters across the different resolution settings for the fastMNN (A) and Seurat (C) integrations. The chosen resolution for the integrations is surrounded in a black box. Proportion plot showing the proportional distribution of the cell types identified in the *RBP6* overexpression and *in vitro* PCF datasets across the clusters of the fastMNN (B) and Seurat (D) integrated object. The black lines on B and D are used to separate the original clusters by whether they are from the Briggs (*in vitro* PCFs) or Laidlaw (*RBP6* overexpression cells) datasets

Throughout this chapter thus far, the assumption has been that the "early_Procyclics" cluster represents cells which are unresponsive to the induced *RBP6* overexpression and, thus, remain as early PCFs. In every integration with the *in vitro* PCFs dataset, apart from fastMNN, the "early_Procyclics" cells cluster mostly by themselves (Fig.35D, Fig.37B, D), indicating that they may not share transcriptional similarity with the *in vitro* PCFs. To test this, the significant DE genes between the "early_Procyclics" cluster and the *in vitro* PCFs were identified. 3235 genes were identified as being significant DE between the "early_Procyclics" cluster and the *in vitro* PCFs, with 3020 genes significantly higher expressed in the *in vitro* PCFs, and 210 genes significantly higher expressed in the "early_Procyclics" cluster (appendix file 28). Given the large amount of genes significantly higher expressed in the *in vitro* PCFs, GO term analysis was carried out on the lists of genes, with a significant GO term defined as benjamini adjusted p-value less than 0.05. Only four significant GO terms were identified for the *in vitro* PCF DE genes, with these being related to RNA and non-coding RNA processing (appendix file 29). 81 significant GO terms were found in the "early_Procyclics" DE genes, with many describing the regulation of cellular metabolism; including the TCA cycle GO term (appendix file 30).

A key experimental difference between the *in vitro* PCFs and the *RBP6* overexpression dataset (beyond the overexpression system) is that the *in vitro* PCFs were grown in SDM-79, while the *RBP6* overexpression cells were grown in SDM-80. The original paper describing SDM-80 (Lamour et al. 2005) showed that levels of glucose consumption decrease and levels of proline consumption increase in SDM-80 grown PCFs. Thus, differences between the two populations may exist in terms of glucose and proline metabolism-associated genes.

For overall differences in metabolism, TCA cycle metabolic pathway genes were significantly enriched in the DE genes associated with the "early_Procyclics" cluster, while no significantly enriched pathways were seen in the *in vitro* PCFs DE genes (appendix 31 & 32). Focusing on specific metabolism genes, the *in vitro* PCFs had significantly higher expression of phosphofructokinase (Tb927.3.3270), phosphoglycerate kinase (Tb927.1.700) and pyruvate kinase 1 (Tb927.10.14140), while the "early_Procyclics" had significantly higher expression of cytochrome oxidase subunit IV (Tb927.1.4100), glutamate dehydrogenase (Tb927.9.5900), citrate synthase (Tb927.10.13430), Fumarate hydratase class I (Tb927.3.4500), glycosomal malat dehydrogenase (Tb927.10.15410) and hexokinase (Tb927.10.2010) (appendix file 28). These DE genes were found in Christiano and colleagues (Christiano et al. 2017) to be higher expressed at the mRNA level in *RBP6* overexpression-derived metacyclics and *in vitro* PCFs, respectively. This pattern did not hold completely, however, with glycosomal glyceraldehyde 3-phosphate dehydrogenase (Tb927.6.4280) being significantly higher expressed in the *in vitro* PCFs scRNA-seq cluster, but was higher expressed in *in vitro* PCFs in the Christiano and colleagues (Christiano et al. 2017) paper. Furthermore, alternative oxidase (Tb927.10.7090) was significantly higher expressed in the "early_Procyclics" cluster but identified as higher expressed in *RBP6* overexpression derived metacyclics in the Christiano and colleagues (Christiano et al. 2017) paper (appendix ref 28). These results fit with the different metabolism requirements of the cell, with the "early_Procyclics" having higher expression of PCF-associated metabolism genes due to the glucose-sparse proline-rich culture media, while the *in vitro* PCFs have more metacyclic-associated metabolism genes, due to the high concentration of glucose in their culture media. These results may suggest that differences in culture environment have caused transcriptomic expression differences between the *in vitro* PCFs and the possible *RBP6* overexpression unresponsive cells.

19.6 Integration consistency of *T. brucei* scRNA-seq datasets

The integrations of the *RBP6* overexpression scRNA-seq dataset with the *in vitro* PCF and *in vivo* insect life cycle stage scRNA-seq datasets showed a lack of consistent results across the three integration methods tested. The common denominator of these integrations was the *RBP6* overexpression dataset, leaving open the possibility that the integration methods simply fail to consistently integrate that particular dataset.

A benchmarking of integration method consistency on *T. brucei* scRNA-seq data was thus performed on the Seurat V5 CCA, Harmony and fastMNN integration methods. The hypothesis for this test is that consistent integrations will be those where the k-nearest neighbours to every cell in the integrated space is similar across all integrations. Inspired by Chari and Pachter (Chari and Pachter 2023), consistency was tested by finding the 30 nearest neighbours of each cell, in the integrated reduced dimension space, and calculating Jaccard similarity (Jaccard 1901) of the neighbours for each cell across the three integrations. An explanation of how Jaccard similarity was used in this context can be seen in figure 16. The lower the Jaccard similarity, the less similar the cells are and the higher the score the more similar they are

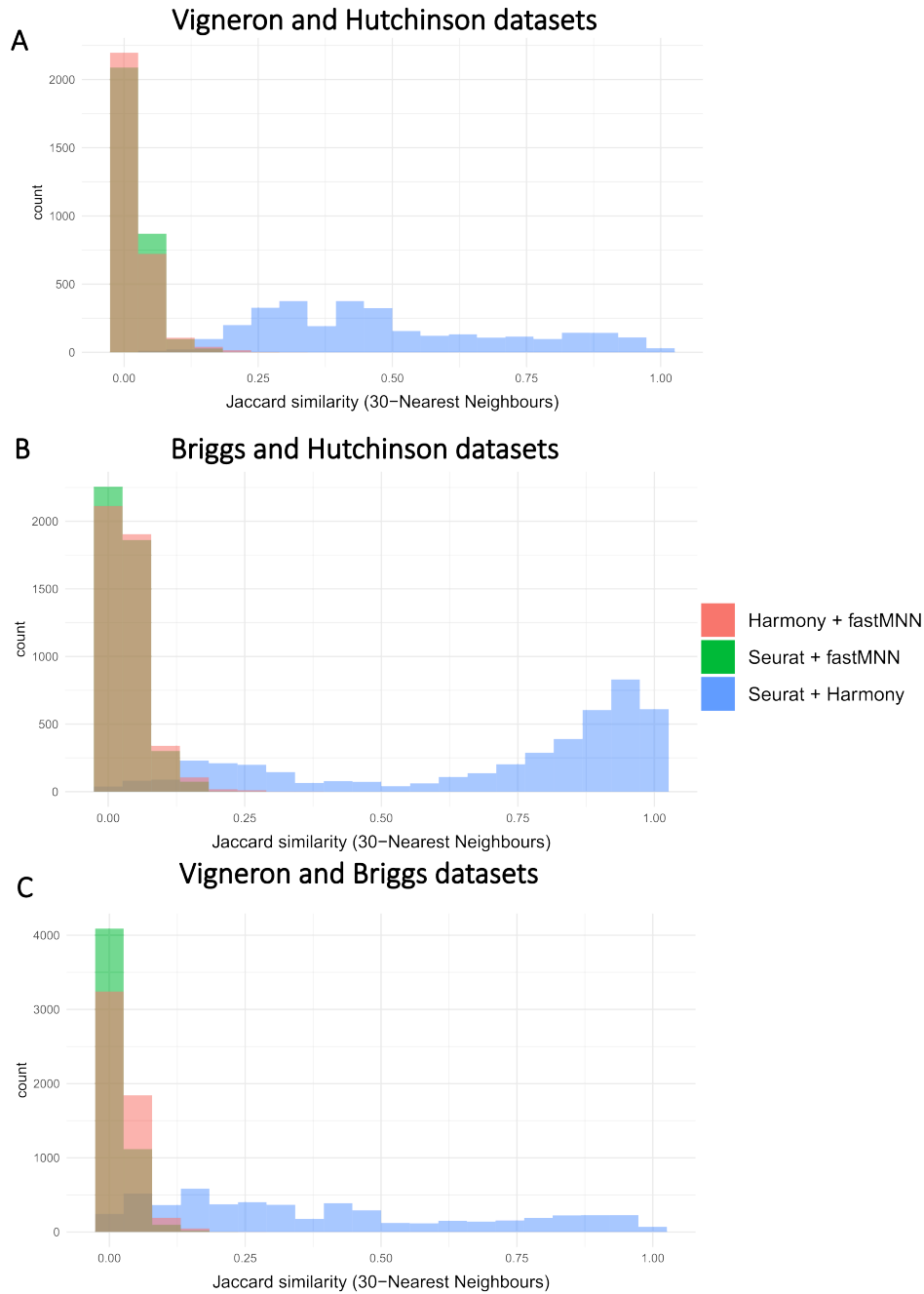


Figure 38: Graphical explanation of Jaccard similarity

Histograms of the Jaccard similarities of a given cells nearest neighbours across different integration pairwise comparisons: Harmony and fastMNN (red), Seurat V5 CCA and fastMNN (green) and Seurat V5 CCA and Harmony (blue). Integrations were down across the Vigneron (10X Chromium *in vivo* insect life cycle stage dataset) and Hutchinson (inDrop *in vivo* insect life cycle stage dataset) (A), Briggs (*in vitro* PCFs) and Hutchinson (B), and Vigneron and Briggs (C). The higher the Jaccard similarity score, the more similar the cell neighbour lists are

To test the consistency of the integration methods, they were applied to the pairwise combinations of the following three scRNA-seq datasets: *in vitro* PCFs (Briggs, Marques, et al. 2023), *in vivo* insect life cycle stage cells captured using inDrop (Hutchinson et al. 2021), and *in vivo* insect life cycle stage cells captured using 10X Chromium (Vigneron et al. 2020). Across all the dataset pairwise integrations, fastMNN had very low Jaccard similarity with either Seurat V5 CCA or Harmony (Fig.38A, B & C). Conversely, Seurat V5 CCA and Harmony Jaccard similarity had more similarity with one another, with a median Jaccard similarity of around 0.75 in the *in vitro* PCFs and inDrop *in vivo* insect life cycle stage cell integration (Fig.38B); however, for the *in vitro* PCFs and 10X Chromium *in vivo* insect life cycle stage cell integration, the median Jaccard similarity was around 0.25 (Fig.38C). These results show a general lack of consistency of the three integration methods, with fastMNN sharing very little overlap in cell neighbours with the other integration methods.

20 Chapter 3: Materials & Methods

20.1 Preparation of PCF *T. brucei* culture media

20.1.1 Procedure to make Semi defined media 79 (SDM-79)

A glass flask containing 450 ml (premade and sterilised) incomplete SDM-79 was put in a 37°C waterbath for 2 minutes. Inside of a fume hood, 50ml of Fetal Bovine Serum (FBS), 1ml of Haemin and 5ml of Streptomycin/penicillin was added to the glass flask of incomplete SDM-79 and was mixed through swirling. Another sterilised glass flask was taken into the fume hood as well as a Millipore (®)Steritop. The Steritop was screwed onto the empty glass flask and a vacuum pipe attached to the Steritop nozzle. A vacuum was created before the complete SDM-79 was poured into the Steritop. The complete, filtered SDM-79 was stored long term in a fridge at 4°C .

20.1.2 Procedure to make Semi defined media 80 (SDM-80)

In a 2 litre plastic container, the components listed in table 5 were added (this table is derived from a protocol written by Dr Federica Giordani):

Table 5: Components and their required volumes/masses for creating Semi defined media 80 (SDM-80)

Component	1 litre	2 litre
NaH ₂ PO ₄	157 mg	314 mg
NaCl	6.8 g	13.6 g
MgSO ₄	100 mg	200 mg
KCl	400 mg	800 mg
CaCl ₂	200 mg	400 mg
L-Arginine	100 mg	200 mg
L-Methionine	70 mg	140 mg
L-Phenylalanine	80 mg	160 mg
L-Threonine	350 mg	700 mg
L-Tyrosine	100 mg	200 mg
Taurine	160 mg	320 mg
L-Alanine	200 mg	400 mg
L-Asparagine	13.2 mg	26.4 mg
L-Aspartate	13.3 mg	26.6 mg
L-Glutamate	14.7 mg	29.4 mg
L-Glutamine	200 mg	400 mg
Glycine	7.5 mg	15 mg
L-Serine	60 mg	120 mg
HEPES	8 g	16 g
MOPS	5 g	10 g
NaHCO ₃	2.2 g	4.4 g
Pyruvate	220 mg	440 mg
Mercaptoethanol (neat)	35 μ l	70 μ l
Hypoxanthine	14 mg	28 mg
Thymidine	4 mg	8 mg
Vitamins 100 X (Sigma, M6895)	10 ml	20 ml
Essential amino acids 50 X (Gibco BRL, 1130-036)	20 ml	40 ml
Phenol Red	4 ml	8 ml
Hemin (2.5 mg/ml) (dissolved in 0.1 M NaOH)	2 ml	4 ml
Dialysed Foetal Calf Serum	100 ml	200 ml
H ₂ O	up to 1 litre	up to 2 litre

The pH of the solution was then adjusted 7.4, with pH measurements carried out with a Fisher Scientific Accumet AE150.

The media was then filtered sterilised and stored at 4°C

The SDM-80 media was then supplemented with 10 mM of Proline before filter sterilisation carried out again.

20.2 *In vitro* culturing of PCF *T. brucei*

29:13 *T. brucei* PCF cells were cultured and maintained in complete, filtered SDM-79 media. The cells were stored in non-vented sterile flasks at 27°C in 5% CO₂.

Drugs were added to the cultures at concentrations detailed in table 6.

Table 6: **Drug concentrations for selecting *T. brucei* PCF cells.** Table showing the final concentrations of each drug

Drug name	Concentration
Hygromycin	50 µg/ml
G418 (Neomycin)	10 µg/ml
Phleomycin	2.5 µg/ml
Puromycin	1 µg/ml

20.2.1 Calculating *T. brucei* culture concentrations

A glass coverslip was placed on a Neubauer chamber haemocytometer and 10 µl of cell culture was pipetted in-between the haemocytometer and the glass coverslip. Cells were counted on the 25 central squares (each square dimensions = 0.2 mm x 0.2 mm), specifically the top left, top right, bottom left, bottom right and central square.

Where the total cell count = (*cellCount* × 5) × (1 × 10⁴)

20.3 Gel electrophoresis

0.5 g of Agarose powder was added to a conical flask, before the addition of 50 ml of 1X Tris-acetate-EDTA (TAE) buffer. The solution was microwaved at 800 watts for 60 seconds, after which the flask was removed (ensuring hands are protected from the heat) and cooled by applying cold water gently to the outside of the flask. The flask was swirled while the cold water was applied to ensure the solution did not start to solidify. 2.5 µl of SYBR Safe DNA gel stain was then added, and the mixture swirled once again to mix the stain into the agarose solution. The solution was then poured into a gel dock, before a gel comb was added. Any bubbles were moved away from the comb towards the end or edge of the gel using a pipette tip. The gel was then allowed to set at room temperature, before the gel comb was removed.

The gel dock was inserted into the gel container and flooded with 1X TAE buffer, ensuring the gel wells were completely flooded and no section of gel was above the surface. For each sample loaded onto the gel, a 10:1 DNA to DNA gel loading dye (ThermoFisher) mixture was created. 50 µl of DNA/loading

dye mixture was then added to the wells, along with 10 μl of Invitrogen 1kb plus DNA ladder. A particular voltage was then applied to the tank for a specific amount of time.

20.4 Gel extraction

The following procedure is adapted from the QIAquick $\text{\textcircled{R}}$ Gel Extraction kit.

Using ultra violet light, the desired band was excised from the gel using a razor. The gel was transferred to an eppendorf and its mass calculated. A gel volume was calculated, with a gel volume being equal to 100 μl 100mg of gel. 3 times the gel volumes of QG buffer was added to the eppendorf which was then incubated at 50 $^{\circ}\text{c}$ for 10 minutes (or until the gel has melted) with vortexing every 2-3 minutes. 1 gel volumes worth of isopropanol was then added, and was mixed through swirling. The contents of the eppendorf were then transferred to a QIAquick spin column, which was then centrifuged 1 minute at 18000g. The flow through was then discarded. 750 μl of PE buffer was then added to the QIAquick column, before further centrifugation at 18000 x g for 1 minute. The flow through was again discarded before the QIAquick column was centrifuged again at 18000g for 1 minute. The QIAquick column was then placed into a microcentrifuge tube (with its cap removed but not thrown away) and 25 μl of EB buffer was added to the centre of the QIAquick column membrane. The sample was left for 5 minutes before centrifugation at 18000g for 1 minute. The sample was stored long term in a -18 $^{\circ}\text{c}$ freezer.

20.5 Creating ampicillin-resistance agar plates

14 g of LB broth+agar powder was added to 400ml of deionized water in a capped glass flask. The cap of the flask was loosened slightly before it was put into a water bath autoclave set at 55 $^{\circ}\text{c}$ for 30 minutes. After 30 minutes, the autoclave water bath was switched off and the mixture was left for a further 20 minutes. The glass flask was then taken out of the autoclave water bath and 400 μl of 100 μg μl concentration Ampicillin was added. The Ampicillin was only added when the glass flask had cooled down, but was still warm. The bench work surface was cleaned with ethanol and a Bunsen burner set up, before around 20ml of the molten agar mixture was added to 20 different petri dishes. The dishes were stored long term in a 5 $^{\circ}\text{c}$ cold room.

20.6 Assessing concentration of DNA

Nucleic acid concentration was assessed using a NanoDrop. Briefly, 1 μl of nucleic acid samples was pipetted onto the NanoDrop pedestal, and the nucleic acid concentration, 260/280 ratio and 260/230 ratio were noted.

20.7 Creating the *RBP6* overexpression plasmid

20.7.1 Amplifying *RBP6* gene from *T. brucei* genomic DNA with polymerase chain reaction

Lister 427 *T. brucei* genomic DNA (gDNA) was acquired from Dr Mark Girasol. The primers used to amplify the *RBP6* sequence were derived from Kolev and colleagues (Kolev, Ramey-Butler, et al. 2012) paper, with the only change being the substitution of an adenine to a thymine on the 25th base of the

Table 7: ***RBP6* gene binding primers.** Sequence of primers used to amplify the *RBP6* gene sequence. The substituted adenine to thymine nucleotide between the Kolev and colleagues (Kolev, Ramey-Butler, et al. 2012) reverse primer and our reverse primer is highlighted in red.

Primer name	Primer sequence (5' - 3')
<i>RBP6</i> forward primer	TATATAAAGCTTATGTTCTACCCCAACAGCCCGCAGACCTCGCCACAGC
<i>RBP6</i> reverse primer	CATCATGGATCCTCAACCAGCGGC T CCGCGGGAACCCGCATGAACG

reverse primer. The primer sequences can be found on table 7.

The polymerase chain reaction (PCR) 50 μ l mastermix components and volumes for amplifying the *RBP6* sequence are as follows: 1 μ l 10 mM dNTP mix (Promega), 10 μ l 5X Phusion® High-Fidelity (HF) Buffer (New England Biolabs; NEB), 0.5 μ l Phusion® HF DNA Polymerase (NEB), 2.5 μ l of *RBP6* forward and reverse primer and 1.18 μ l of Lister 427 gDNA

A negative control mastermix was also made, with the volume of gDNA being replaced with deionised water. A mastermix was also made with the inclusion of 1.5 μ l Dimethylsulfoxide (DMSO), which changes the total volume of deionised water added to 30.82 μ l.

The PCR was performed using a MJ Research PTC-200 Peltier thermal cycler with the following sequence of temperatures and lengths: 1. 98°C for 30 seconds, 2. 98°C for 10 seconds, 3. 72°C for 1 minute, 4. Repeat step 2 thirty times, 5. 72 °c for 10 minutes. The PCR mix was then stored at 4°C until removed for long term storage at -4°C.

Gel electrophoresis of the PCR product was carried out, with 100 volts passed through the gel for 60 minutes. A gel extraction was then carried out on the *RBP6* PCR product.

20.7.2 Extracting pLEW100v5 plasmid backbone

The pLEW100v5 plasmid was ordered from AddGene.

The following restriction enzyme double digest protocol was obtained from NEBcloner. A double restriction enzyme digest was carried out on the pLEW100v5 plasmid with high fidelity (HF) HindIII and BamHI restriction enzymes. In an eppendorf tube, 1 μ g of pLEW100v5 plasmid DNA, 5 μ l of 10X rCutSmart buffer, 0.75 μ l of HF HindII and HF BamHI restriction enzymes was added, and made up to 50 μ l with deionised water. The eppendorf was incubated at 37°C for 15 minutes before transfer to ice. Gel electrophoresis was performed on the digested pLEW100v5 plasmid solution at 100 volts for 60 minutes. The backbone of the pLEW100v5 gel band (~5000bp) was removed and put through gel extraction. The extracted pLEW100v5 backbone was then stored long term at -18°C.

20.7.3 Ligation of *RBP6* gene sequence into pLEW100v5 plasmid

The *RBP6* gene PCR product was digested in a similar fashion to the pLEW100v5 plasmid (see the previous sub sub section).

The following protocol for ligation of enzyme was adapted from NEBcloner T4 DNA Ligase (M0202) protocol. A ligation mix was created with the following components and volumes: 2 μ l of T4 DNA ligase

buffer, 1 μ l of T4 DNA ligase, 3.1 μ l of digested *RBP6* sequence (\sim 68ng of DNA), 4 μ l of pLEW100v5 backbone (\sim 50.8 ng of DNA) and 9.9 μ l of deionised water. The ligation mix was mixed gently using a pipette tip and left at room temperature for 10 minutes. The enzymes were heat inactivated on a 65 $^{\circ}$ C heatblock for 10 minutes, before transfer to ice for 5 minutes. The solution was then immediately used for transformation into DH5- α *E. coli*. This plasmid will henceforth be referred to as OE-plasmid

20.7.4 Transformation of engineered *RBP6* overexpression plasmid into DH5- α *E. coli*

The protocol for transforming the engineered OE-plasmid into DH5- α *E. coli* was adapted from the NEBcloner High Efficiency Transformation Protocol (C2987H). DH5- α *E. coli* were thawed on ice before the entire volume of ligated OE-plasmid was added. The tube was flicked 5 times to mix cells. The mixture was placed on ice for 45 minutes before it was heat shocked at 42 $^{\circ}$ C for 30 seconds. The sample was placed on ice for 5 minutes, before the mixture was transferred to a tube of room temperature SOC media and placed in a shaker for 60 minutes at 37 $^{\circ}$ C.

50 μ l of transformed bacteria were transferred onto the ampicillin Agar plates and spread, before the plates were incubated at 37 $^{\circ}$ C overnight (\sim 12-18 hours).

20.7.5 Extraction and purification of engineered *RBP6* overexpression plasmid from DH5- α *E. coli*

Colonies from the plates were picked with a pipette tip and placed into a falcon tube of liquid LB, then incubated at 37 $^{\circ}$ C overnight (\sim 12-18 hours).

The following procedure is taken from the QIAprep $\text{\textcircled{R}}$ Spin Miniprep kit protocol. 2.5 ml of the overnight bacterial culture was centrifuged at 6800 g for 3 minutes at room temperature. The pelleted bacteria was then resuspended in 250 μ l of QIAprep P1 buffer before transfer to a microcentrifuge tube. 250 μ l of QIAprep P2 buffer was then added and the tube mixed through inversion 4-6 times, or until the solution became clear. This step should not be allowed to run for no more than 5 minutes. 350 μ l of QIAprep N3 buffer was then added to the tube before mixing through inverting the tube 4-6 times. The tube was then centrifuged for 10 minutes at 18000g.

The supernatant of the tube was then poured into a QIAprep spin column, before centrifugation for 60 seconds at 18000g. The flow-through was discarded, before 0.75 ml of QIAprep PE buffer was added to the QIAprep spin column. The spin column was then centrifuged again for 1 minute at 18000 g, and the flow-through again discarded, before another centrifugation step was carried out for 1 minute at 18000 g.

The QIAprep column was then placed into a microcentrifuge tube with its cap broken off (but still kept). 25 μ l of QIAprep EB buffer was then added to the center of the spin column and left for 5 minutes before centrifugation at 18,000 x g for 1 minute. The resulting purified OE-plasmid was stored long term in a freezer at -18 $^{\circ}$ C.

Gel electrophoresis was performed on the miniprep extracted sample. 1 g of agarose and 10 ml TAE was used to create the gel, and 100 volts was passed through the gel for 54 minutes.

20.7.6 Sanger sequencing of *RBP6* overexpression plasmid insert

Sanger sequencing was carried out using the Eurofins mix2seq kit and service. Briefly, 1.54 μl of purified *RBP6*-pLEW100v5 plasmid, 13.46 μl of deionized water, and 2 μl of either JM42 or HDK430 primer 8 were added to the mix2seq kit tubes and were sent off for Sanger sequencing.

The two resulting Sanger sequences were aligned (individually) against the sequence of the Lister 427 *RBP6* gene (Tb427.03.2930) using EMBOSS Needle (Madeira et al. 2024) and Benchling.

Table 8: Sequence of Sanger sequence primers used to amplify the *RBP6* gene insert from the pLEW100v5 plasmid.

Primer name	Primer sequence (5' - 3')
JM42	TTGAAGACTTCAATTACACC
HDK430	TAACCAACCTGCAGGCG

20.7.7 Transfection of engineered *RBP6* overexpression plasmid into 29:13 *T. brucei* cells

The following procedure is adapted from the NEBcloner NotI-HF restriction enzyme digest protocol (Order R3189). 10 μg of OE-plasmid was linearised using the NotI HF restriction enzyme. The master mix was set up with the following components and volumes: 50 $\mu\text{g}/\text{ml}$ of linearised OE-plasmid, 50 μl of 10X rCutSmart buffer, 10 μl of HF-NotI restriction enzyme. The solution was made up to 500 μl with deionised water. The mastermix was then incubated at 37°C for 10 minutes before long term storage in a -18°C fridge.

Transfection buffer (Burkard, Jutzi, and Roditi 2011) was created with the following molar concentrations dissolved in deionised water: 90mM sodium phosphate, 5 mM potassium chloride, 0.15 mM calcium chloride, 50 mM HEPES. The solution was filter sterilised, before long term storage at 4°C.

2×10^7 29:13 Lister 427 PCFs cells (maintained in SDM-79) were collected into a 50ml falcon tube and centrifuged for 5 minutes at 1000g. The supernatant was removed and the cells were resuspended in 200 μl of transfection buffer, before being centrifuged for a further 5 minutes at 1000 g. The supernatant was discarded and 200 μl of transfection buffer was added slowly, gently resuspending the cells. 10 μg of linearised OEplasmid was then added and the suspension mixed gently. The suspension was then transferred into a BioRad 0.2 cm electrode gap cuvette and placed in an Amaxa electroporator, before being zapped on the X-0001 program. The cells were then transferred to a flask containing SDM-79 media with no drugs, and incubated overnight at 27°C.

The next day, 9 ml of cell culture was added to a 50 ml falcon tube and centrifuged at 1000 x g for 10 minutes. While the cells were spinning, conditioned media was made with the following components: 6 ml of FBS, 9 ml of media isolated from active cell culture and 45 ml of fresh culture. 2.5 $\mu\text{g}/\text{ml}$ of Phleomycin was also added to the media to select for cells which have successfully had the OE-plasmid transfected.

Three 50 ml falcon tubes were filled with a 1:10, 1:100 and 1:1000 ratio of transfection culture to conditioned media. The dilutions were then transferred to a plastic reservoir and 180 μl diluted culture was added to each well of a 96 well plate. The plates were then sealed with parafilm and incubated for a week and a half at 27°C. The plates were monitored for growth in the wells, with any wells that had

cell growth being isolated and cultured in SDM-79 with selection drugs, before long term storage at -80°C.

20.8 Inducing overexpression of *RBP6* in 29:13 *T. brucei* cells

RBP6 OE 29:13 PCFs were maintained in either SDM-79 or SDM-80 at a concentration of 2×10^5 cells/ml. The cells were induced each day with 10 $\mu\text{g}/\text{ml}$ of tetracycline and the required drugs to maintain an *in vitro* PCF culture.

20.9 Immunofluorescence assay of *RBP6* overexpressing *T. brucei* PCF cells

An anti-CRD antibody was sourced from Dr Lucy Glover at the Institut Pasteur.

A blocking solution of 1% Bovine serum albumin, 0.2% Tween and 1X PBS was created.

Dilutions of antibody in blocking solution were made as specified in table 9.

Table 9: Dilutions of antibody

Antibody	Dilution
anti-CRD rabbit	1:5000
anti-EP Procyclin mouse	1:750
594 anti-rabbit	1:1000
488 anti-mouse	1:1000

A five day *RBP6* overexpression inducement was carried out, as previously described, on the *RBP6* overexpression cells gifted by Dr Lucy Glover. The cells were maintained in SDM-80 during the course of the inducement. At the same time, a culture of uninduced cells were also maintained (in SDM-79). Samples were taken at days three, four and five post-inducement for the induced and uninduced cultures. Table 10 details the four slides that were made up for each day.

The following steps occur within hood. 2×10^6 cells were collected and transferred to a falcon tube. The cells were then centrifuged at 1,500 rpm for 10 minutes, before the supernatant was extracted. The cells were then resuspended in 1 ml of PBS and transferred to an eppendorf. The cells were centrifuged again at 3000 rpm for three minutes. $\sim 50 \mu\text{l}$ of supernatant was left in the eppendorf while the rest was removed. The pellet was resuspended in the remaining supernatant.

Table 10: Slide makeup for IFA imaging for each day of the inducement sampled

Slide No.	Antibodies	Condition	DAPI
1	Yes	Induced	Yes
2	No	Induced	Yes
3	Yes	Uninduced	Yes
4	No	Uninduced	Yes

200 μl of poly-L-lysine was pipetted onto lens tissue and the surface of a slide with reaction wells was wiped with it a few times. Using a PAP pen, a circle was drawn around the reaction wells which were to have cells in them.

The $\sim 50 \mu\text{l}$ of resuspended cells was then pipetted into one of the wells and left to incubate for 5 minutes. The supernatant was then removed carefully and slowly via pipetting, before 25 μl of 4% formaldehyde was added to the well and left for 4 minutes. The formaldehyde was then removed carefully and slowly via pipetting, before 50 μl of glycine was added to the slide and left for 10 minutes. The glycine was then removed carefully and slowly via pipetting, and the glycine addition step was repeated once more. 50 μl of PBS was then added to the well and left for 5 minutes, before it was carefully and slowly removed via pipetting. This step was repeated once more.

25 μl of the blocking solution was added to the well and left for 1 hour inside a wet chamber (a container with a wet paper towel inside). The blocking solution was removed slowly and carefully via pipetting.

If antibody was to be added to the well, 25 μl of primary antibody was added and allowed to incubate inside the wet chamber for 1 hour. The primary antibody was removed slowly and carefully via pipetting before 50 μl of PBS was added to the well and left for 5 minutes, then removed. The PBS wash was repeated once more. 25 μl of secondary antibody was added and allowed to incubate inside the wet chamber for 1 hour. The primary antibody was removed slowly and carefully via pipetting before 50 μl of PBS was added to the well and left for 5 minutes, then removed. The PBS wash was repeated once more.

5 μl of DAPI was then added to the well and allowed to incubate in the wet chamber for 4 minutes. A cover slip was then added, and any air bubbles were squeezed out. The coverslip and slide were sealed using nail varnish, and the slide was stored in a opaque box at 4°C long term.

20.10 Imaging of *RBP6* overexpression IFA slides

Z-stack images were taken on a Leica DiM8 microscope using the DAPI, GFP and Y5 channels. The images were processed using Fiji (Schindelin et al. 2012).

20.11 Single cell RNA sequencing of *RBP6* overexpressing 29:13 *T. brucei* cells

The following steps outlining the procedure for inducing *RBP6* overexpression in PCFs, was adapted from a protocol from Dr Lucy Glover.

The *RBP6* overexpressing PCF cells were transferred to a flask containing SDM-79 and allowed to recover for two days. The culture (referred to hereafter as the "control" flask) was then diluted down to 2×10^6 cells/ml and 1 ml of this culture was transferred into an eppendorf, before being centrifuged at 800 g for 5 minutes. The supernatant was then removed by pipetting and the cells resuspended in 1 ml of SDM-80. The resuspended cells were then transferred into a culture flask marked "Day 5" and selection drugs were added at their respective concentrations, as well as 10 μg of tetracycline per ml of culture volume. This process was repeated for the next two days, with the creation of a Day 4 culture flask the next day and a Day 3 culture flask the day after that.

At the start of each day, all culture flasks were diluted down to 2×10^6 cells/ml and a selection drugs added based on the volume of media added to dilute the cells. The control flask was diluted with SDM-79 and no tetracycline was added, while the Day 5, 4 and 3 flasks were diluted with SDM-80, and $10 \mu\text{g}$ of tetracycline was added per ml of media to dilute the cells.

While the inducement experiment was occurring, a PSG + 0.04% bovine serum albumin (BSA) solution was created. The PSG solution consisted of (for 500 ml of PSG): 0.244 g of sodium phosphate monobasic dihydrate, 1.275 g of sodium chloride, 4.04 g of disodium phosphate, 75 g of D-glucose and 500 ml of deionised water. The pH of the solution was then adjusted to 7.8 using hydrochloric acid.

The following steps outlining the procedure for preparing *T. brucei* for 10X scRNA-seq was adapted from a protocol from Dr Emma Marie Briggs.

After 5 days from the start of the experiment, the cell concentrations were all calculated and 2×10^6 cells from the Day 5, Day 4 and Day 3 culture flasks were transferred to separate eppendorf tubes, before being centrifuged for 10 minutes at 400 g. The supernatant was removed via pipetting with wide bore pipette tips, and the cells resuspended in 1 ml of PSG + 0.04% BSA solution with wide bore pipette tips. This centrifugation and resuspension was repeated twice, but on the last repeat only $200 \mu\text{l}$ of PSG + 0.04% BSA was added to resuspend the cells. The cells were then passed through a Flowmi® cell strainer and the concentration of cells in each sample was calculated.

The Day 4 and Day 5 samples were all diluted down to a concentration of 1.2×10^4 cells/ml and the Day 3 sample diluted down to 6×10^3 cells/ml using PSG and 0.04% BSA. $10 \mu\text{l}$ of Day 4 and Day 5 and $6 \mu\text{l}$ of Day 3 were added to the same eppendorf tube and mixed via stirring with a pipette tip. $17 \mu\text{l}$ of the sample mix was then put into an eppendorf on ice and transferred to undergo the 10X sequencing process, which was carried out by Julie Galbraith of University of Glasgow Polyomics, using the 10X 3' V3 kit.

20.12 Mapping the *RBP6* overexpression single cell RNA-seq data

The combined TREU927 with kinetoplastid DNA (kDNA) genes genome reference, with the 3' untranslated regions extended by 2500 base pairs, was obtained from Dr Emma Marie Briggs (Briggs, Rojas, et al. 2021). The metacyclic VSG (mVSG) gene sequences were taken from Hutchinson and colleagues (Hutchinson et al. 2021) (who in turn obtained them from Müller and colleagues (Müller et al. 2018)) and appended to the 3' UTR extended combined reference. The scRNA-seq data was mapped against the combined TREU927 UTR extended + kDNA + mVSG reference using Cellranger Count (version 7.0.0) with default settings.

20.13 Analysis of the *RBP6* overexpression dataset

20.13.1 Quality control and preprocessing

The filtered count matrix generated from the 10X Cellranger count pipeline was loaded into R as a Seurat object.

Genes which were expressed in less than 10 cells were filtered out of object, and cells which had a total gene count of less than 400 or more than 2500 were removed. The percentage of reads which map

to ribosomal RNA (rRNA) or kinetoplastid RNA (kRNA) was calculated for each cell and cells whose percentage rRNA was more than 4.5% and kRNA was more than 5% were removed from the object. The expression data was then normalised using Seurat NormalizeData() function, with the scale factor set as the median of total UMI counts across the cells. The top 2000 most variably expressed genes were identified with Seurat and used as the feature space basis of the dataset.

The normalised expression was then scaled to have a mean of 0 and a standard deviation of 1, and the effect of total UMI count was regressed out. The scaled, regressed expression data was then used as the basis for performing Principal Component Analysis (PCA), with the first 50 principal components (PCs) calculated. The standard deviation of the 50 PCs was calculated and used to determine the PC cutoff. A k-nearest neighbors (kNN) graph was then constructed from the cell embeddings in the first 11 PCs, with the edges of the cell connections refined using Jaccard Similarity (Jaccard 1901) of each cells nearest neighbours. This refined nearest neighbours graph was used as the basis to perform Louvain clustering (Blondel et al. 2008). Louvain clustering was performed at resolutions of: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0 and the results were visualised using Clustree (Zappia and Oshlack 2018). Using the Clustree output, a final clustering resolution of 0.6 was chosen. The first 12 PCs were then used as the basis for performing Uniform Manifold Approximation and Projection (UMAP), with the 12 PC space being reduced down to a 2D UMAP (McInnes et al. 2018).

20.13.2 Cell cycle labelling of the *RBP6* overexpression scRNA-seq dataset

Each cell in the dataset was assigned a cell cycle phase label of: early G1, late G1, S or G2-Mitosis phase using the method outlined in Briggs and colleagues (Briggs, Marques, et al. 2023). Briefly, a list of genes associated with each of the four phases (derived from the Briggs and colleagues paper (Briggs, Marques, et al. 2023)) was used to create a meta feature of cell cycle phase using Seurat's MetaFeature function. The meta feature scores were then normalised by dividing by the mean score, and phase labels were assigned based on which cell cycle phase had the highest score. If multiple meta feature scores were above 1.05 the cell was assigned a value of "undecided" and if all were less than 0.9 then they were assigned a value of "unlabelled".

20.13.3 Analysis of the *RBP6* overexpression process

The analysis described in the previous paragraph was repeated twice, with the only changes being that one repeat had cell cycle genes removed from the highly variable genes list and the second repeat being that, along with total UMI count, the four phases meta feature score were regressed out. The rest of the analysis described in the subsection involves the data generated when regressing out the total UMI count and the four phases meta feature scores.

Marker genes for the cell cycle and UMI regressed dataset were defined as genes whose log₂ fold change (log₂FC) was greater than 0.25, expressed in at least 5% of cells in either the target cluster or the rest of the dataset and had a Bonferroni corrected p-value of less than 0.05.

All of the genes of the TREU927 genome were extracted from TriTrypDB (Shanmugasundram et al. 2023), and the genes whose product description contained "ribosomal protein" or "ribosomal subunit protein" were extracted. This list of genes was then filtered to only include genes that were expressed in at least 10% of cells in the dataset. A ribosomal protein expression meta feature was then created, using the filtered gene list, by Seurat's MetaFeature function. The meta feature of a cell was then normalised

by the mean of all the meta feature values for the cells. Pearson correlation was then carried out on the relationship between the ribosomal protein meta feature, and the PC values of the first four PCs and a p-value calculated using linear regression, with the alternative hypothesis being that there is a significant negative correlation between the two variables. Significance for this test was set at a p-value of less than 0.05.

A new analysis was performed with cell cycle genes and ribosomal protein genes removed from the highly variable genes and total UMI count, the four phases meta feature score and the ribosomal protein feature score regressed out. Clustering and dimensionality reduction was completed as set out previously, with the number of PCs used being 15 and the clustering resolution being 0.5.

20.13.4 Comparison of scRNA-seq and bulk RNA-seq *RBP6* overexpression data

The raw count matrices of the induced *RBP6* overexpression (and uninduced control) bulk RNA-seq samples were taken from Doleželová and colleagues (Dolezelova et al. 2020). The count matrices had their genes transformed into the syntenic orthologs of the TREU927 genome, before being processed with DESeq2 with default parameters.

The top 10 marker genes (in terms of \log_2FC) were extracted for each cluster and their scaled normalized expression was plotted for the bulk RNA-seq datasets. As some genes in the top 10 marker gene list were not present in the bulk RNA-seq data, the list of top marker genes (in terms of \log_2FC) was searched for the next most highly expressed gene until 10 genes were returned.

20.13.5 Cluster annotation of *RBP6* overexpression scRNA-seq data

Clusters were annotated based on the following criteria. Cluster 0 was annotated as "ZC3H36_EP_hi_GPEET_mid" as a significant marker of the cluster was the expression of *ZC3H36*, and it had high expression of the EP procyclins and middle expression of GPEET procyclin. Cluster 1 was annotated as "RHS_EP_hi_GPEET_mid" as several of its markers were Retrotransposons, and it had high expression of the EP procyclins and middle expression of GPEET procyclin. Cluster 2 was annotated as "histone_EP_hi_GPEET_mid" as several of its markers were histones, and it had high expression of the EP procyclins and middle expression of GPEET procyclin. Cluster 3 was annotated as "EP_hi_GPEET_mid" as it had high expression of the EP procyclins, and middle expression of GPEET procyclin. Cluster 4 was annotated as "late_RBP6_PAG" because the top 10 marker genes of that cluster (in terms of \log_2FC) were highly expressed in the final days of the *RBP6* overexpression bulk RNA-seq data, and several of its markers were PAGs. Cluster 5 was annotated as "early_Procyclics" because the top 10 marker genes of that cluster (in terms of \log_2FC) were highly expressed in the early days of the *RBP6* overexpression bulk RNA-seq dataset, and it highly expressed GPEET procyclin. Cluster 6 was annotated as "late_Metacyclics" because the top 10 marker genes of that cluster (in terms of \log_2FC) were highly expressed in the final days of the *RBP6* overexpression bulk RNA-seq data, and marker genes of the cluster included several VSG pseudogenes and a metacyclic form marker gene *SGM1.7*.

20.14 Integration of *RBP6* overexpression scRNA-seq data and *in vivo* insect stage scRNA-seq data

The integrated rds object was taken from the Hutchinson and colleagues (Hutchinson et al. 2021) paper, and the "Lib1" batch cells (i.e. the cells from the first sequencing run performed by the author) isolated. Cell cycle phases were assigned as previously described, as well as the ribosomal protein meta feature score calculated. Metabolic pathway enrichment analysis was carried out with KEGG and Metacyc (Caspi et al. 2014 & Kanehisa et al. 2017)

The Hutchinson *in vivo* dataset and the *RBP6* overexpression dataset were then merged together, with only the 7,131 genes that were present in both the original datasets kept. The datasets were normalised using Seurat, with the scale factor being the median of the total UMI across the cells in both datasets. The top 2000 variable features were then identified with Seurat, and the normalised expression values of the merged dataset were scaled. The effect of UMI, cell cycle phase meta feature score and ribosomal protein meta feature score was regressed out. The data were then integrated together using the Harmony (Korsunsky et al. 2019) method of integration. The Harmony embeddings were then used as the basis for Louvain clustering, and UMAP embeddings were also generated from the Harmony embeddings, with the first 12 Harmony embedding dimensions used. A clustering resolution of 0.3 was chosen based on the Clustree, generating 8 clusters. The above steps were repeated using Seurat V5 Canonical Correlation Analysis (CCA) (Y. Hao, T. Stuart, et al. 2024) and fastMNN (Haghverdi et al. 2018) integration, with the number of integrated dimensions used being the first 12 and 15 and the clustering resolutions being 0.8 and 0.5, respectively.

DE gene analysis was carried out between the "Metacyclics" cluster of the Hutchinson *in vivo* insect stage dataset, and the "late_possibleMetacyclics" cluster of the *RBP6* overexpression dataset. Significantly DE genes were identified as genes with Bonferroni corrected p-value < 0.05, absolute $\log_2FC > 0.5$, and were expressed in at least 10% of cells in one of the clusters being compared.

20.15 Integration of *RBP6* overexpression scRNA-seq data and *in vitro* PCFs

The integrated PCF rds object was taken from the Briggs and colleagues (Briggs, Marques, et al. 2023), paper and the fresh log-phase PCF cells were kept in the dataset. The Lister 427 2018 genome gene IDs of the dataset were transformed into their syntenic orthologs in the TREU927 genome. The Briggs PCF dataset and the *RBP6* overexpression dataset were then merged together, with only the 7013 genes that were present in both the original datasets kept. The datasets were normalised using Seurat, with the scale factor being the median of the total UMI across the cells, in both datasets. The top 2000 variable features were then identified with Seurat, before the normalised expression values of the merged dataset was scaled and the effect of UMI was regressed out. The data was then integrated together using the Harmony (Korsunsky et al. 2019) method of integration. The Harmony embeddings were then used as the basis for Louvain clustering and UMAP embeddings were also generated from the Harmony embeddings, with the first 11 Harmony embedding dimensions used. A clustering resolution of 0.4 was chosen based on the Clustree, generating 7 clusters. The above steps were repeated using Seurat V5 Canonical Correlation Analysis (CCA) (Y. Hao, T. Stuart, et al. 2024) and fastMNN (Haghverdi et al. 2018) integration, with the number of integrated dimensions used being the first 11, and the clustering resolutions being 0.6 and 0.4, respectively. Metabolic pathway enrichment analysis was carried out with KEGG and Metacyc (Caspi et al. 2014 & Kanehisa et al. 2017)

20.16 Testing consistency of integration methods on *T. brucei* scRNA-seq data

The fastq files from NCBI BioProject ID PRJNA562204 (Vigneron et al. 2020) were downloaded and mapped against the combined TREU927, 2500bp 3' UTR extension plus kDNA genes genome reference using Cellranger Count (v7.0.0). Quality control of the cells was carried out as specified by the original paper.

Pairwise integrations of the *in vitro* PCF (Briggs, Marques, et al. 2023), inDrop *in vivo* insect life cycle stage cells (Hutchinson et al. 2021) and 10X Chromium *in vivo* insect life cycle stage cells (Vigneron et al. 2020) scRNA-seq datasets was carried out as follows. The datasets were normalised with the Seurat NormalizeData function, with the scale factor being the median of total UMIs across the two datasets. The top 2000 variable features were then identified with Seurat, before the normalised expression values of the merged dataset was scaled and the effect of UMI was regressed out. PCA was then run on the scaled regressed expression values, with the first 50 PCs being captured. The datasets were then integrated with Seurat V5 CCA, fastMNN and Harmony separately, and the 30 nearest neighbours for each cell in the integrated space was calculated using the kNN function from dbSCAN (Ester et al. 1996). Jaccard similarity was then assessed using the Jaccard function from mlr3measures.

20.17 Package Versions

The analyses described in this chapter were carried out using R version 4.3.2.

Specific package versions are as follows: Seurat 5.0.1, harmony 1.2.0, dbSCAN 1.2-0, clustree 0.5.1, batchelor 1.19.1, stringr 1.5.1, presto 1.0.0, DESeq2 1.42.0, ggplot2 3.5.2, mlr3measures 0.5.0

21 Chapter 3: Discussion

In this chapter an implementation of the *RBP6* overexpression model of the insect stage life cycle of *T. brucei* was attempted, and scRNA-seq was carried out on a working example of the *RBP6* overexpression model. The results of this analysis are broadly inconclusive as to what processes drive the transition of the parasite through its insect life cycle stages.

The *RBP6* gene was first amplified from *T. brucei* Lister 427 gDNA and inserted into a tetracycline inducible overexpression construct. The engineered construct was successfully used to generate two overexpression cell lines. However, sequencing revealed several unexpected differences in the *RBP6* coding region compared to that of the reference genome. As this resulted in several changes to the expressed RBP6 protein compared to the published model (Kolev, Ramey-Butler, et al. 2012), these cell lines were not continued with. A validated model of *RBP6* overexpression was thus sourced from Dr Lucy Glover of the Institut Pasteur.

In the uninduced day three samples, all cells were fluorescent for EP procyclin, as expected. No metacyclics are expected at this timepoint, yet many cells also expressed CRD. Metacyclics lose EP procyclin surface protein expression in favour of VSG (Acosta-Serrano et al. 2001), meaning the IFA results do not fit with established biology. These results are inconsistent with the scRNA-seq data, which showed very few cells expressing mVSG. Thus the CRD IFA may not be optimal, lacking either antibody specificity or incorrect IFA procedure.

The seeming lack of epimastigotes in both the microscopy analysis and scRNA-seq results is an interesting finding. With respect to the microscopy, there are several possible reasons for this absence. The first is that as an antibody marker of epimastigotes (i.e. BARP) was not available, second is that the kinetoplastid and nucleus localisation does not appear to be a 100% determinant of epimastigotes in the *RBP6* overexpression system. Doleželová and colleagues (Dolezelova et al. 2020) performed IFAs using BARP on *RBP6* overexpressing cells, and found cells which express surface BARP, but do not have the correct kinetoplastid and nucleus positioning for epimastigotes. This disparity in the kinetoplastid and nucleus positioning was the key reason why the distances between them were not interpreted. The IFA results and lack of evidence of epimastigotes in the scRNA-seq data mean that whether epimastigotes are generated by the *RBP6* overexpression process cannot be determined.

At day five post-*RBP6* overexpression, all cells that were visualised were fluorescent in the anti-CRD channel. Previous accounts from Dr Lucy Glover (personal communications) indicated that sampling days three, four and five post *RBP6* overexpression induction would lead to the generation of some, but not many, metacyclics, and only at around day five would they begin to appear. Therefore, while it is more accurate that all the cells are fluorescent in anti-CRD channel at day five post-inducement than day three post-inducement, these results still do not reflect previous results of this implementation of the *RBP6* overexpression system and indicate that non-specific fluorescence was detected with anti-CRD staining. The localisation of anti-EP procyclin on the surface of the day five post-inducement samples presents one of the most interesting results of the IFA analysis. No previous literature has described the "spot" localised surface expression of EP procyclin on *T. brucei* cells. The images, however, look similar to intracellular imaging of the protein, where EP procyclin cannot be trafficked to the surface, and instead it localises near the Golgi apparatus (L. Liu et al. 2013). Assuming that the day five post-induction results represents intracellular EP procyclin, these results could represent the EP procyclin trafficking to the surface being switched off and the switching on of mVSG trafficking. However, it should be noted that no permeabilisation of the cells was carried out during the IFA procedure, and thus intracellular protein should not be detectable.

Taken at face value, as all the uninduced day three and induced day five sample cells displayed, by fluorescence, CRD expression, it indicates that *RBP6* overexpression had occurred in both the samples. As the uninduced should not be overexpressing *RBP6*, the results may indicate that overexpression of *RBP6* is leaky. In both the IFA and scRNA-seq growth curves, the uninduced samples undergo drops in cell concentration, which could be reflective of *RBP6* overexpression occurring, lending more weight to the hypothesis that the cell line has leaky *RBP6* overexpression.

Although the image analysis could not conclusively identify what life cycle stages were present in the data, the transcriptomic output of the scRNA-seq data allows some insight into this aspect.

Combining the results of the *in vivo* insect life cycle stages and *in vitro* culture PCF integrations with the *RBP6* overexpression dataset, it suggests that the "late_possibleMetacyclics" cluster cells are neither *RBP6* overexpression unresponsive PCFs, nor are they fully formed metacyclics. For the former, this was initially suggested through the decreased expression of *GPEET* compared with the *in vitro* PCFs. As *GPEET* is a marker of *in vitro* PCFs, if the "late_possibleMetacyclics" cluster did represent *RBP6* overexpression unresponsive PCFs, one would expect similarly high expression of *GPEET*, but this is only seen in the "early_Procyclics". Furthermore, while the *in vitro* PCFs were cultured in SDM-79 and the *RBP6* overexpression cells were cultured in SDM-80, the latter should rely more heavily on proline for its energy. Despite this, proline catabolism genes *PDH* and *DP5CDH* were more lowly expressed in the "late_possibleMetacyclics" cluster cells compared with the *in vitro* culture PCFs, suggesting that these cells have started to undergo a metabolic switch away from proline as a major energy source.

Evidence supporting the conclusion that the "late_possibleMetacyclics" cluster does not represent fully formed metacyclics is mainly driven by the lack of mVSG and *SGM1.7* expression in most of the cells. Although the fact that some of these cells do have high expression of these genes suggests some of the cells have become fully formed metacyclics, or that the switching on of mVSG happens early on in the differentiation process. When compared to *in vivo* metacyclics, the expression of metacyclic marker genes is lower in the *RBP6* overexpression "late_possibleMetacyclics" cluster and the expression of PCF marker genes are higher. However, as the expression of the PCF marker genes are not as high as in the *in vitro* PCFs, it may suggest these cells are transitioning towards the metacyclic stage, with a few of the cells reaching the metacyclic stage, as seen by high mVSG expression.

The most likely reasons for why so few metacyclic forms were seen is that a late enough time point, post-*RBP6* overexpression induction, was not captured. Observations by Dr Lucy Glover found that metacyclics begin to appear at day 5 post-*RBP6* overexpression induction, but do not peak at this time point. The reason a later time point was not taken is that, while metacyclics increase in proportion after day five post-inducement, the amount of cell death also increases. As these dead cells could end up captured in the scRNA-seq data, later time points were not taken. The reasons for why cell death is higher towards the end of the *RBP6* overexpression process is not known. It could possibly be that prolonged induced *RBP6* overexpression is lethal in cells, or that the concentration of glucose in SDM-80 is not high enough to sustain the metacyclics, which have changed their metabolism to utilise glucose as the major energy source.

Discussing the prospect of whether the "early_Procyclics" cluster does in fact represent *RBP6* overexpression unresponsive PCFs, the evidence in this chapter suggests they might be, but this is not confirmed. The cluster has *GPEET* high and EP procyclin mid expression, indicative of early PCFs (Knüsel and Roditi 2013), and the marker genes of this cluster are highest expressed in the uninduced bulk RNA-seq datasets of *RBP6* overexpression cells, giving evidence towards the conclusion that they are *RBP6* overexpression unresponsive cells. The lack of integration of the "early_Procyclics" with the *in vitro* PCFs is the most direct evidence against the conclusion that the "early_Procyclics" represent *RBP6* overexpression unresponsive cells, as the *in vitro* PCFs should, in effect, be similar to PCFs that do not respond to *RBP6* overexpression, i.e. static in terms of life cycle stage. Furthermore, over 3000 genes were significantly DE between the groups, representing around 27% of all the genes annotated in the TREU927 genome (Berriman et al. 2005) and around 37% of the 8589 genes (that passed the quality control threshold) captured in the *RBP6* overexpression scRNA-seq dataset. There are thus two competing points of evidence: the comparison with the *RBP6* overexpression bulk RNA-seq data suggesting that the "early_Procyclics" cluster cells are *RBP6* overexpression unresponsive cells, and the *in vitro* PCF scRNA-seq data suggesting they are not. In my opinion, the evidence that the marker genes of the "early_Procyclics" cluster are most highly expressed in the *RBP6* overexpression uninduced bulk RNA-seq sample is more compelling than the transcriptomic differences between "early_Procyclics" cluster and the *in vitro* PCF cells. This is primarily motivated by the fact the culture environments of the *RBP6* overexpression bulk RNA-seq and scRNA-seq cells were the same, i.e. the cells were cultured in SDM-80. While metabolomic changes induced by SDM-79 and SDM-80 have been assessed (Lamour et al. 2005), no paper has identified the transcriptomic or proteomic changes that they cause in PCFs. This makes it difficult to know whether the gene expression changes that are seen between the "early_Procyclics" and the *in vitro* PCFs reflect genuine changes caused by *RBP6* overexpression, or whether this reflects transcriptome changes brought about by differences in energy sources.

With PCFs and (albeit few) metacyclics confirmed in the transcriptomic data, the "elephant in the room" from the scRNA-seq data is the absence of epimastigotes, which should be generated by the *RBP6*

overexpression system. The expression of epimastigote marker *BARP* was consistently sparse across all the cells after *RBP6* overexpression, with mRNA expression levels never reaching the peaks they do in *in vivo* derived epimastigotes. As some metacyclic-like cells were generated in the scRNA-seq dataset, it stands to reason that epimastigotes should have been generated, as they are the intermediary stage between PCF and metacyclics. There are several possible reasons for why epimastigotes did not appear in the scRNA-seq data. The first, and perhaps most unlikely, is that the *RBP6* gene being overexpressed is the mutated form described in Shi and colleagues (H. Shi, K. Butler, and Tschudi 2018). As discussed in the introduction, this model fails to generate epimastigotes, as no *BARP* protein expression was detectable in the samples. This conclusion is unlikely for a variety of reasons. The *RBP6* gene sequence in the OE-plasmid was checked by Dr Lucy Glover, and the mutation was not identified in the OE-plasmid insert. The likelihood of this mutation happening spontaneously in culture, and then becoming the dominant form of PCF is thus very low. Interestingly, while *BARP* protein expression is absent in the mutated *RBP6* overexpression cells, mRNA expression of *BARP* is still detectable (H. Shi, K. Butler, and Tschudi 2018).

Another possible explanation for the lack of epimastigotes is that no life cycle stages beyond PCFs were present in the day three, four or five post-*RBP6* overexpression induction, i.e. no epimastigotes and no metacyclics were present. While this explanation is more likely than the previous, I would argue it is still unlikely. The most direct evidence against this is the presence of cells whose marker genes are highest expressed at the end points of the *RBP6* overexpression bulk RNA-seq data, and cells which highly express mVSG can be identified. Ultimately, no satisfying answer can be found for why an epimastigote transcriptome signature is not present in the dataset. At present, I can think of no probable reason why cells with epimastigote transcriptomic signatures were not identified, thus it appears to be a quirk of this particular scRNA-seq dataset, as it is not a signature that is seen in other transcriptomic analyses of the system (Dolezelova et al. 2020).

Due to the uncertainty in terms of life cycle stage allocation (and the apparent lack of others) to many of the cells in the *RBP6* overexpression dataset, future single cell transcriptomic work with this model may benefit from the use of plate-based, rather than droplet-based, methods. In the Malaria Cell Atlas paper (Howick, Russell, et al. 2019), the authors utilisation of Smart-seq2 (Picelli et al. 2014) meant that they could isolate the life cycle stages and assign them to wells, allowing them to know the identity of the cell without having to rely on the cells' transcriptome. In terms of the *RBP6* overexpression cells, it could be possible to isolate them based on their characteristics, such as kinetoplastid and nucleus positioning, or certain genes (like *BARP* or *PAG*) could be tagged with mCherry to allow the sorting of cells before the transcriptome is captured, helping eliminate some of the uncertainty when it comes to identifying life cycle stages.

Thus far, a lot of emphasis has been placed on the analysis of the "late_possibleMetacyclics" and the "early_Procyclics" clusters. This is due to the fact they represent clusters which have the most noticeable similarity with defined life cycle stages of the parasite or have a strong association with a particular timepoint of the *RBP6* overexpression. This next section of discussion shall thus focus on the other clusters, whose place in the *RBP6* overexpression process (or indeed the *in vivo* life cycle) is not clear.

The "ZC3H36_EP_hi_GPEET_mid" cluster was defined as such due to its *GPEET* and *EP* procyclins expression levels and the presence of *ZC3H36* (Tb927.10.12760) in its marker gene list, although this gene is also highly expressed in the "late_possibleMetacyclics" cluster. The expression of this gene, coupled with the lack of epimastigotes in the data, gives more weight to the hypothesis that the mutated *RBP6* overexpression system has been captured, as it is uniquely upregulated in the mutated *RBP6* overexpression form compared to the uninduced cells, with the same upregulation not seen when comparing the WT

model with uninduced cells (H. Shi, K. Butler, and Tschudi 2018). This picture is not clear cut, as only two of the other similarly upregulated transcripts were identified as significant marker genes of the clusters, with both of these two being marker genes of the "early_Procyclics" cluster, which are hypothesised to be *RBP6* overexpression uninduced cells, thus they don't share the same pattern as that described in the mutated *RBP6* overexpression model paper (H. Shi, K. Butler, and Tschudi 2018). Within parasites across the tsetse fly tissue, *ZC3H36* transcript is highest expressed in those in the proventriculus (Savage et al. 2016 & Naguleswaran et al. 2021), where PCFs, mesocyclic PCFs and epimastigote forms are present (Sharma et al. 2009). No direct investigation of *ZC3H36* has been carried out, so its functions are not understood. Although, given the importance of RNA-binding proteins in allowing life cycle transitions (Kolev, Ramey-Butler, et al. 2012, Mugo and Clayton 2017, Briggs, Rojas, et al. 2021, B. Liu, Kamanyi Marucha, and Clayton 2020 & Cayla et al. 2020), it may play a role in the transitions or processes that occur in the proventriculus. It is thus possible that the "ZC3H36_EP_hi_GPEET_mid" cluster may represent cells undergoing such a transition in the *RBP6* overexpression model. Complicating this picture is the high expression of *ZC3H36* in the "late_possibleMetacyclics" cluster, a population that *in vivo* would be found in the salivary glands, where expression of *ZC3H36* mRNA is low (Naguleswaran et al. 2021).

The RHS proteins are split into 7 subfamilies (RHS1-RHS7) across 118 genes, with 78 of these genes being pseudogenes (Florini et al. 2019). The functions of the subfamilies vary. RNAi of RHS6 causes an accumulation of cells lacking a nucleus and associates with several replication factor C subunit proteins, suggesting a role in DNA replication (Florini et al. 2019), RHS4 is part of the RNA polymerase II complex (Devaux et al. 2006, Das et al. 2006), suggesting a role in transcription, and RHS2 is found to immunoprecipitate with ribosomal and RNA-binding proteins (Florini et al. 2019), suggesting a role in post-transcription processes. Therefore, expression of RHS encoding genes in the *RBP6* overexpression cells may be related to the fact that the parasites will have to undergo many morphological and physiological changes due to the effect of *RBP6* overexpression, and RHS helps facilitate these changes. This hypothesis is strengthened by the fact that many of the RHS are lowest expressed in the uninduced samples from the bulk RNA-seq *RBP6* overexpression dataset.

In both the Kolev and Doleželová papers (Kolev, Ramey-Butler, et al. 2012 & Dolezelova et al. 2020), life cycle stage proportion changes across the *RBP6* overexpression inducement. The start of the process is dominated by PCFs, before giving rise to epimastigotes, then to metacyclics and then, at the very end, PCFs begin to become more prevalent again. Kolev and colleagues (Kolev, Ramey-Butler, et al. 2012) did not comment on this trajectory; however Doleželová and colleagues (Dolezelova et al. 2020) offer two explanations for this observation. 1: the cells are dying once the metacyclic stages are reached and the PCFs left are those that are unresponsive to *RBP6* overexpression and thus replicate and become the main proportion of cells in the culture. 2: the cells revert back to PCFs, after metacyclics stages have been achieved.

As discussed, the early (possibly *RBP6* overexpression unresponsive) PCFs identified in the *RBP6* overexpression scRNA-seq data and the PCFs from the *in vitro* scRNA-seq dataset (Briggs, Marques, et al. 2023) are marked by high *GPEET* expression. If the PCFs that appear at the end of the overexpression timecourse were *RBP6* overexpression unresponsive PCFs, there should be an upregulation of *GPEET* expression, consistent with more early PCFs, at the end of the *RBP6* overexpression inducement time course. Doleželová and colleagues (Dolezelova et al. 2020), however, show that *GPEET* expression does not increase in the last timepoint (where PCFs make up 50% of cells); in fact, across the entirety of the *RBP6* overexpression-induced samples it remains downregulated in comparison to uninduced cells. This observation suggests that the PCFs at the end of the *RBP6* overexpression are responsive in some way to the *RBP6* overexpression process (or the environment it creates in culture), as there are key

transcriptomic differences between the uninduced and the late stages of *RBP6* overexpression. In the scRNA-seq *RBP6* overexpression dataset, there are two clusters annotated as being "late" in the *RBP6* overexpression process, as their marker genes were highly expressed in the late stages of *RBP6* overexpression in the bulk RNA-seq data. One of these clusters ("late_RBP6_PAG"), had higher EP procyclins expression and lower *GPEET* expression than the "early_Procyclics", which could indicate that they are later stage PCFs (Knüsel and Roditi 2013). Thus, this cluster could represent the PCF-like cells that are seen at the end of the *RBP6* overexpression process.

The main markers of "late_RBP6_PAG" cluster are the PAGs, named as such as they are located downstream of the procyclins and contained within the same transcription unit (Koenig-Martin, Yamage, and Roditi 1992), with *PAG1*, *PAG2* and *PAG3* increasing in terms of mRNA expression across BSF to PCF differentiation (Haenni et al. 2006). The functions of these genes are not known, although they are not essential for cell survival or infectivity (Haenni et al. 2006). In the insect stages, mRNA expression of the PAGs is high in the midgut but low in the salivary glands (Naguleswaran et al. 2021), making them higher expressed in PCFs than in metacyclics/epimastigotes, which would fit with our conclusion that the cluster that expresses them has PCF-like gene expression. These results could imply that *RBP6* overexpression leads to the generation of PCFs that are not transcriptionally similar to uninduced PCFs. Whether or not this is due to the direct effects of *RBP6* or some environmental signal is not known, nor is it known what the direct descendants of these cells are in terms of life cycle stage. It is, however, an interesting point of future enquiry to try and identify the lineage and cause of these cells.

While it can be expected that the cell cycle would have an influence on the transcriptome of the cells (and thus their clustering and reduced dimension embeddings), the finding that ribosomal protein gene expression also has a strong influence is interesting. There is an argument to be made that ribosomal protein gene expression should not have been regressed out as it may represent an important marker of the *T. brucei* insect life cycle stages. While this is a valid point, previous scRNA-seq papers that have investigated the single cell and bulk transcriptomic changes that occur across the insect stages (Vigneron et al. 2020, Naguleswaran et al. 2021, Hutchinson et al. 2021 & Howick, L. Peacock, et al. 2022), show that a variety of non-ribosomal protein genes can be used to separate out the major insect life cycle stages, therefore regressing out ribosomal protein gene expression may not negatively impact the results. There is a paucity of research on ribosomal protein gene expression in *T. brucei*, but some conclusions can be drawn from the literature that does exist. As mentioned, mRNA stability is an important mechanism for controlling gene expression in *T. brucei*. Ribosomal protein encoding genes are some of the most stable transcripts in *T. brucei*, with their half life often exceeding 120 mins (Manful, Fadda, and Clayton 2011) and the transcripts being more stable in PCFs compared with mammalian forms (Fadda et al. 2014), suggesting some life cycle stage specific regulation.

While the Hutchinson, Naguleswaran and Vigneron papers (Vigneron et al. 2020, Hutchinson et al. 2021 & Naguleswaran et al. 2021) do not comment on ribosomal protein gene expression changes across the *in vivo*, Howick and colleagues (Howick, L. Peacock, et al. 2022) find that a variety of genes which contribute towards ribosomal GO terms are highest expressed in the middle time points of salivary gland colonisation and lowest at the later timepoints, suggesting that it could play a time dependent role in salivary gland colonisation, and thus possibly metacyclic cell development. Furthermore, Christiano and colleagues (Christiano et al. 2017) found that 106 out of the 200 highest expressed genes in PCFs, compared with *RBP6* overexpression derived metacyclics, encoded ribosomal proteins. As ribosomal protein gene expression appears to be associated with different life cycle stages of *T. brucei*, whether or not this is a driver of life cycle stage development or a byproduct represents an area of biology which should be studied more.

Accuracy of integrations was not tested, only consistency. As discussed in the thesis general introduction, assessing biological accuracy of integrations can be fraught with issues where the extent of cell heterogeneity is not appreciated, and thus, assessing integration accuracy would require an entire chapter in of itself. As such, no certain comment can be made on which integration method is the best for *T. brucei*. A suggestion from this chapter might be that fastMNN performs well, as the integrations of the *RBP6* overexpression cells with the *in vitro* PCFs and *in vivo* insect life cycle stages was, seemingly, the most biologically accurate. What is clear from the results is that there was relatively little agreement over the three integration methods as to the nearest neighbours of the cells. Interestingly, fastMNN had the lowest overall agreement with the other integration methods, but seemed to perform the best on integrating the *RBP6* overexpression cells. Harmony and fastMNN are considered as two of the best performing integration methods for mammalian datasets (Luecken, Buttner, et al. 2022), thus their inconsistent integrations for the *T. brucei* samples show that perhaps their similarly high performance on mammalian datasets does not translate well into *T. brucei*. Given these disparities in integration consistency (and perhaps integration accuracy), research should be dedicated to testing more integration methods on the *T. brucei* scRNA-seq datasets with a focus on biological accuracy of the integration.

If the experiments in this chapter were to be repeated again, there are several aspects that would be changed. The first is that uninduced cells should have been included in the scRNA-seq sample to serve as a comparison for the *RBP6* overexpressed cells. While there is some confidence in saying that the "early_Procyclics" represent *RBP6* overexpression unresponsive cells (and thus practically speaking uninduced cells), more assured confirmation would have come from including uninduced cells in the sample. The second change would be ensuring that the cells are only cultured in one type of media for the length of the experiment. For the IFA and scRNA-seq experiments, the seeding culture of uninduced cells for the day three, four and five post-*RBP6* overexpression inducement samples were maintained in SDM-79. While this means that the cells had three days to adjust to the environment before sampling, the seeding culture should have been maintained in SDM-80 to ensure that the parasites were fully adapted to the environment before *RBP6* overexpression was induced.

The aim of this chapter, initially, was to utilise the *RBP6* overexpression model to identify genes important in transition of *T. brucei* through its insect life cycle stages. Instead, this chapter has become a dissection of the *RBP6* overexpression model itself and how well it represents the *in vivo* insect life cycle stages. If the results of this chapter are taken at face value, it suggests that the *RBP6* overexpression model poorly represents the *in vivo* process due to the lack of transcriptomic signature for key life cycle stages and presence of cells with a transcriptomic that is not seen in the *in vivo* data, i.e. the PAG expressing cells. These results, however, should not be taken at face value for several reasons. First, the results are often contradictory of each other, for example the IFAs suggesting many metacyclic cells were present while the scRNA-seq analysis suggesting only a few were generated. Second, the lack of transcriptomic signature epimastigotes has not been commented on in other papers, albeit no other paper has used scRNA-seq to analyse the *RBP6* overexpression system.

All the papers that have used this model have started from the assumption that the model captures the distinct developmental stages of the parasite in the tsetse fly *in vivo* life cycle of the parasite (H. Shi, K. Butler, and Tschudi 2018, (Dolezelova et al. 2020, H. Shi, K. Butler, and Tschudi 2018, Kolev, Ramey-Butler, et al. 2012, Hutchinson et al. 2021 & Christiano et al. 2017). Given the brevity of the original paper (Kolev, Ramey-Butler, et al. 2012), more work should be done to validate that this is the case, as results within this chapter and other papers (Dolezelova et al. 2020) suggest that it may not be as reflective of the *in vivo* life cycle as the authors present.

22 Conclusion and future directions

Throughout this thesis, a variety of scRNA-seq analysis methods have been applied (and one created) to extract information on the life cycle stages (and transitions between them) of Trypanosomatids.

In chapter one I created TrAGEDy, a method which can accurately align both mammalian stage *T. brucei* and mammalian cell trajectories across different conditions, and reveal gene expression differences between them, which are not revealed by other contemporary methods. One of the reasons why TrAGEDy was developed was to provide a tool which could compare the *RBP6* overexpression induced life cycle with the *in vivo* life cycle of *T. brucei*. This comparison was not carried out due to the unexpected results seen in the *RBP6* overexpression data, such as the lack of pure metacyclic cells and epimastigotes, and the presence of possible unidentified life cycle stages which may be specific to the *RBP6* overexpression model i.e. the PAG expressing PCFs. The application of TrAGEDy to both a mammalian and parasite scRNA-seq dataset was an intentional choice. As stated in the introduction, many scRNA-seq analysis methods are only tested on mammalian datasets, specifically those of human and mouse. As parasites typically have less RNA content than mammalian cells (and in the case of kinetoplastids have a different type of gene expression (Clayton 2019)), these represent differences which could lead to inconsistencies in scRNA-seq analysis method performance. It is for this reason that TrAGEDy was applied to both parasite and mammalian datasets, with the results showing that TrAGEDy performs well in the context of both human and *T. brucei* datasets.

While version one of TrAGEDy outperformed its competitors in terms of alignment and gene expression changes identified, there are several changes which would be included in future versions of the package to address the limitations with the current version. The first limitation is that TrAGEDy can only compare two conditions at a time which, given the availability of perturb-seq (Dixit et al. 2016), leaves it limited in what it can compare. Future versions of the package could thus allow multi-condition comparisons and alignments. Secondly, modern scRNA-seq methods do not integrate or align datasets in a pairwise manor, instead performing these actions for all the datasets at the same time. TrAGEDy requires users to manually align each replicate of a condition, which can be time consuming when many replicates are present in the analysis, so future versions should include ways to perform parallel alignments, not pairwise. Thirdly, TrAGEDy is limited to aligning linear trajectories, which makes its applications to more complex processes, like malaria sexual commitment or haematopoiesis, problematic. Finally, TrAGEDy does not allow additional modality information to be utilised when calculating similarity, so future versions could allow proteomic data, for example, to be included when aligning datasets.

For the second chapter (as part of a collaboration with the University of Dundee) we document the transcriptome of the four major life cycle stages of *T. cruzi* at the single cell level. With this, we create the first scRNA-seq atlas of the four major *T. cruzi* life cycle stages, and show that the atlas can be utilised to accurately annotate novel data, reinforcing its use for the wider *T. cruzi* community by allowing them to quickly annotate their scRNA-seq datasets. Beyond the novelty of the dataset, we capture the process of metacyclogenesis at the single cell level, identifying a variety of novel RBPs which may influence the parasites progression from epimastigote to metacyclic trypomastigote form cells. Finally, I identify possible trypomastigote subpopulations which may have differing roles when it comes to sustaining *T. cruzi* infection in the mammalian host.

There is an implication that atlases represent a definitive database of a biological system. That is not to say it is a ground truth in terms of understanding all the processes that occur within the system, but certainly in terms of the cell types which appear within it. It is my opinion that many atlases (especially mammalian ones) fail in this task, and the reason for this lies in the lack of importance placed on in-depth cell type annotation in many scRNA-seq papers, and the narrowness of markers used to separate cell

types. Referring back to the integration benchmark paper by Luecken and colleagues (Luecken, Buttner, et al. 2022) discussed in the introduction, they defined the difference between NK and NK T cells as being that they both express NKG7 and GZMA, while NK cells lack expression of CD4 or CD8. NKG7, GZMA, CD4/CD8 positive is a narrow definition of what an NK T cell is, and could also describe a conventional CD8+ T cell (Ng et al. 2020). As cell type annotation was not the primary task of the paper they, understandably, did not delve deep into this facet. For atlases, however, correct cell type annotation is perhaps one of the most important parts of the process, but it is often neglected. For example, both the Tabula Muris and Tabula Sapiens extensively profile a variety of organs across mice and humans, respectively. However, they do not annotate any cells as $\gamma\delta$ T cells, a population which is enriched at mucosal sites (G. Q. Li et al. 2023), many of which have been sampled in these atlases.

To the credit of the Tabula Muris authors, the paper contains a lengthy supplementary table detailing marker gene expression across all the different samples, along with their cell type annotations. However, they often define cell types using only one or two genes, which may explain why the $\gamma\delta$ T cells were not identified. The Tabula Sapiens paper opted for automatic annotation of their clusters using onClass (S. Wang et al. 2021), which may have annotated them as NK cells or NK T cells, as they are all cytotoxic leukocyte populations. As $\gamma\delta$ T cells are a well-defined immune subset, whose distribution throughout mouse/human tissues is known, their apparent absence in the data should have been spotted and commented on. All that said does not mean I am blind to the fact that the complexity of entire mammals is beyond that of parasites, and thus this presents a challenge when it comes to analysing and annotating atlas level mammalian datasets. My point, rather, is that there are aspects from parasite atlas papers which could be utilised to give more confidence when annotating. An example of this comes from Howick and colleagues (Howick, Russell, et al. 2019), who utilised SMART-seq for the Malaria Cell Atlas initially, allowing individual life cycle stage cells to be isolated accurately and labelled as such in the output. Following this, they generated 10X scRNA-seq data (allowing them to capture more cells) and used the previous SMART-seq data to provide accurate annotations of the 10X scRNA-seq data.

In the third and final chapter, an attempt was made to utilise the *RBP6* overexpression model of the *T. brucei* insect stage development, in order to capture these life cycle transitions *in vitro*. Instead, the results suggested that this system is perhaps not reflective of the parasites *in vivo* development, as cells with an epimastigote transcriptomic signature could not be identified. The scRNA-seq and IFA data revealed metacyclic-like cells but these did not represent metacyclic cells generated *in vivo*, leaving open the question of whether the *RBP6* overexpression process generates true metacyclics. In the *RBP6* overexpression bulk RNA-seq data from Doleželová and colleagues (Dolezelova et al. 2020), PAG expression is highest at the end of the *RBP6* overexpression process. In the scRNA-seq data generated for this chapter, I show that these PAG expressing cells have the transcriptomic signature of PCFs, giving evidence that these represent the PCF cells present at the end of the *RBP6* overexpression process (Kolev, Ramey-Butler, et al. 2012. Dolezelova et al. 2020) Thus, a point of future interest is elucidating their role in the process, and whether or not these represent a life cycle stage which is found *in vivo*. An interesting finding was how variable the methods were in their results. In the *RBP6* overexpression integrations, Seurat V5 CCA and Harmony were quite similar to one another in terms of their integration, but for the *T. cruzi* atlas they returned very different results. The exact reasoning behind this finding is not known. One hypothesis could be that there exists a diverse range of batch-effect-driven gene expression changes that some integration methods are better suited to correct for than others. It would therefore be useful to the field to understand the reasons why integration methods can be so similar for some datasets but diverge in others. Furthermore, as integration methods are usually tested only on mammalian data, a benchmark study of integration methods on parasite data would be beneficial to the community, identifying methods which may work well for parasite data, but not for mammalian.

The resolution of scRNA-seq makes it an excellent tool for investigating the life cycle stages and

transitions of Trypanosomatids. Using scRNA-seq, many hypothesises have been generated throughout this thesis concerning the validity of the *RBP6* overexpression model of *T. brucei* development, the functional characteristics of *T. cruzi* trypomastigotes and drivers of metacyclogenesis in *T. cruzi*. Future work should be dedicated to confirming these results in the wet lab. This thesis also highlights the lack of attention paid to non-mammalian datasets when it comes to scRNA-seq method development, and highlights some inconsistencies in the performance of well-utilised integration methods across parasite scRNA-seq datasets. This thesis stands on a bedrock of foundational scRNA-seq analysis of kinetoplastids, further contributing to this burgeoning field, and it further shows the power of scRNA-seq to reveal novel findings about the transcriptome of Trypanosomatids.

23 Code Availability

The code used to generate the results of the chapters can be found at the following GitHub repositories:

Chapter 1 - <https://github.com/No2Ross/TrAGEDy>

Chapter 2 - <https://github.com/No2Ross/TcruziAtlas>

Chapter 3 - <https://github.com/No2Ross/TbruceiRBP6>

24 Appendices

24.1 Appendix files

The appendix files (and their descriptions) can be found at the following GitHub repository:

<https://github.com/No2Ross/ThesisAppendixFiles>

24.2 Appendix tables

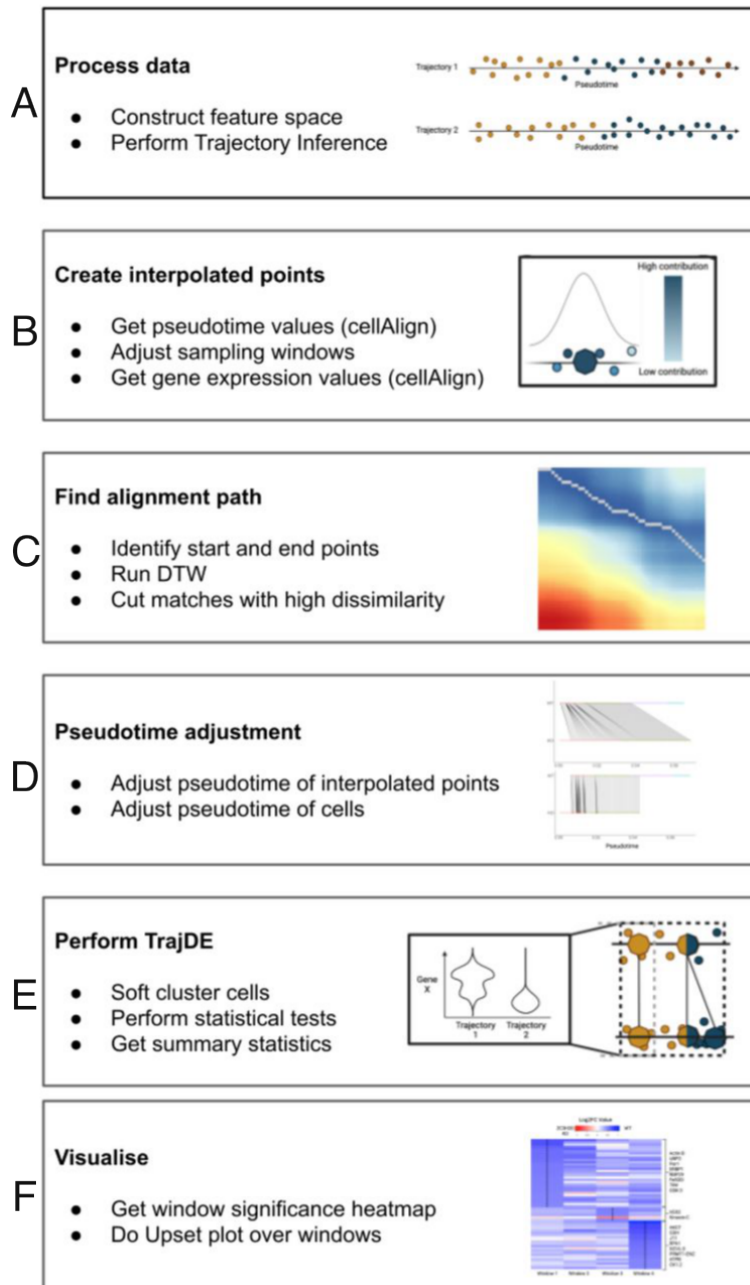
Sample	Gene counts minimum	Gene counts maximum	Percentage reads mapping to kinetoplastid maximum	Percentage reads mapping to human transcriptome maximum
AMA _{6/24}	200	750	20	50
EP/AMA ₁₂₀	400	2000	30	10
ADT	200	1500	25	NA
SE/MT 1	150	1000	40	NA
SE/MT 2	200	1250	40	NA
MIX 1	150	1500	28	30
MIX 2	200	1250	30	30

Appendix Table 1: **Filtering parameters for each of the *T. cruzi* scRNA-seq datasets**

Sample	Number of input reads	Uniquely mapped reads %	% of reads mapped to multiple loci	% of reads mapped to too many loci	% of reads unmapped: too many mismatches	% of reads unmapped: too short	% of reads unmapped: other
bAMA ₆ 1	23,475,929	61.52	15.50	9.83	0	11.24	1.91
bAMA ₆ 2	20,219,803	61.24	17.02	10.92	0	8.76	2.06
bAMA ₂₄ 1	20,498,527	63.24	13.80	5.10	0	16.23	1.62
bAMA ₂₄ 2	24,492,333	63.34	13.64	4.53	0	16.73	1.76
bAMA ₁₂₀ 1	26,077,005	56.16	17.90	16.22	0	6.22	3.50
bAMA ₁₂₀ 2	27,749,110	53.25	17.05	21.31	0	5.44	2.94

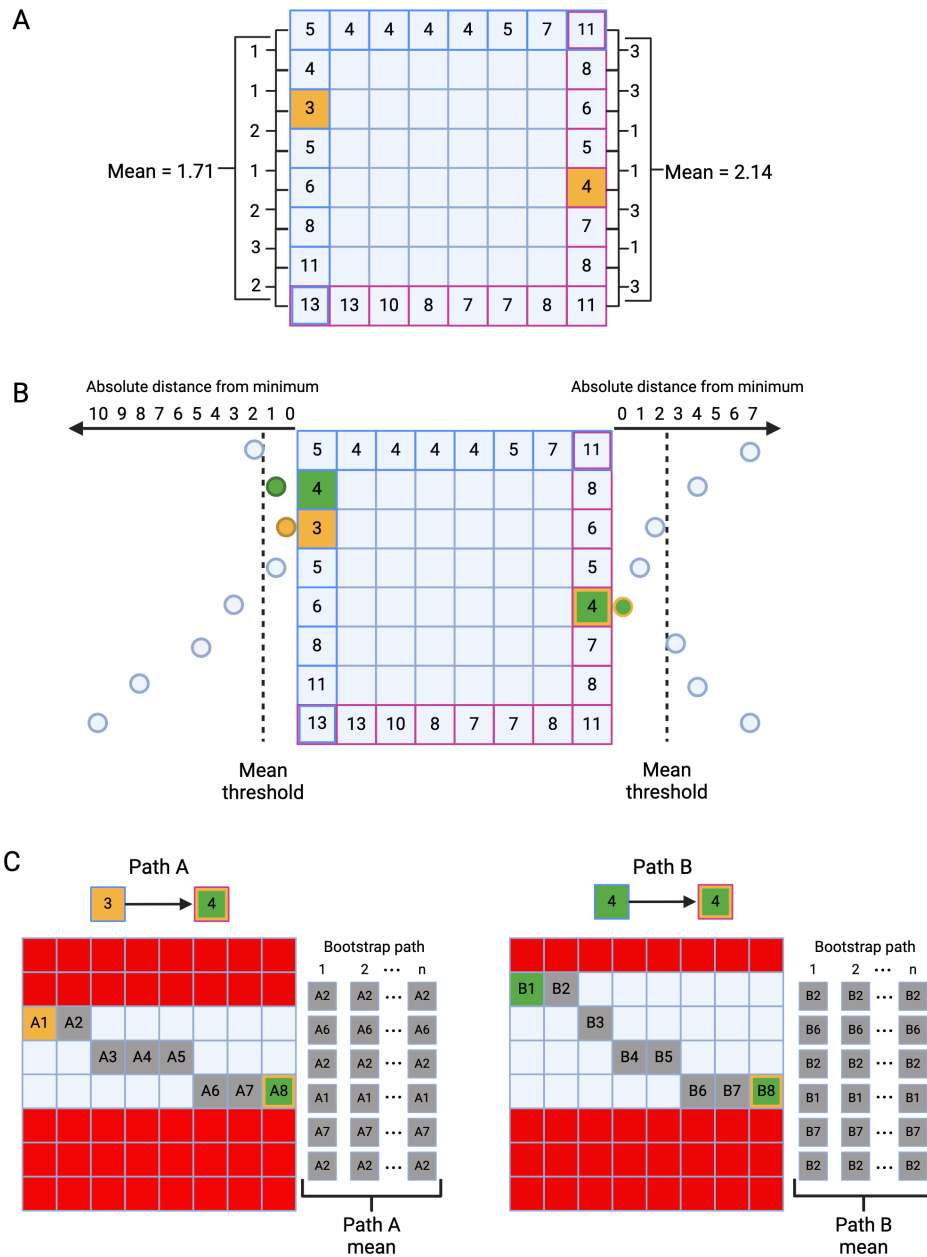
Appendix Table 2: Mapping statistics returned for the *T. cruzi* amastigotes bulk RNA-seq datasets mapped against the Dm28c 2018 + kDNA reference + GRCh38 reference

24.3 Appendix figures



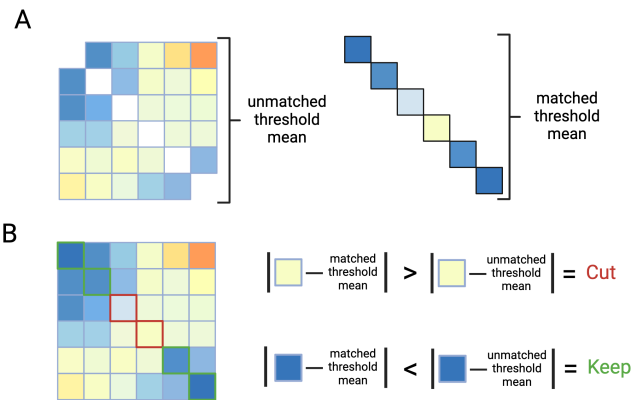
Appendix Figure 1: Workflow of TrAGEDy process.

Workflow detailing the different parts of the TrAGEDy process. Sections which were derived from cellAlign are noted in the process boxes. TrAGEDy starts with the user performing Trajectory Inference (TI) on the datasets individually using a shared feature space (A). The pseudotime and gene expression of the cells is then used to create interpolated points using the method defined by cellAlign (B). From the interpolated points, TrAGEDy then works out the optimal alignment by first identifying the optimal start and end points before running DTW and pruning any suboptimal matches (C). The alignment path is then used to adjust the pseudotime of the interpolated points, and then subsequently the cells themselves (D). DE genes are then identified across the shared process to see identify what gene expression changes occur before the datasets diverge from one another (E). TrAGEDy offers a variety of ways to visualise the changes across the alignment (F).



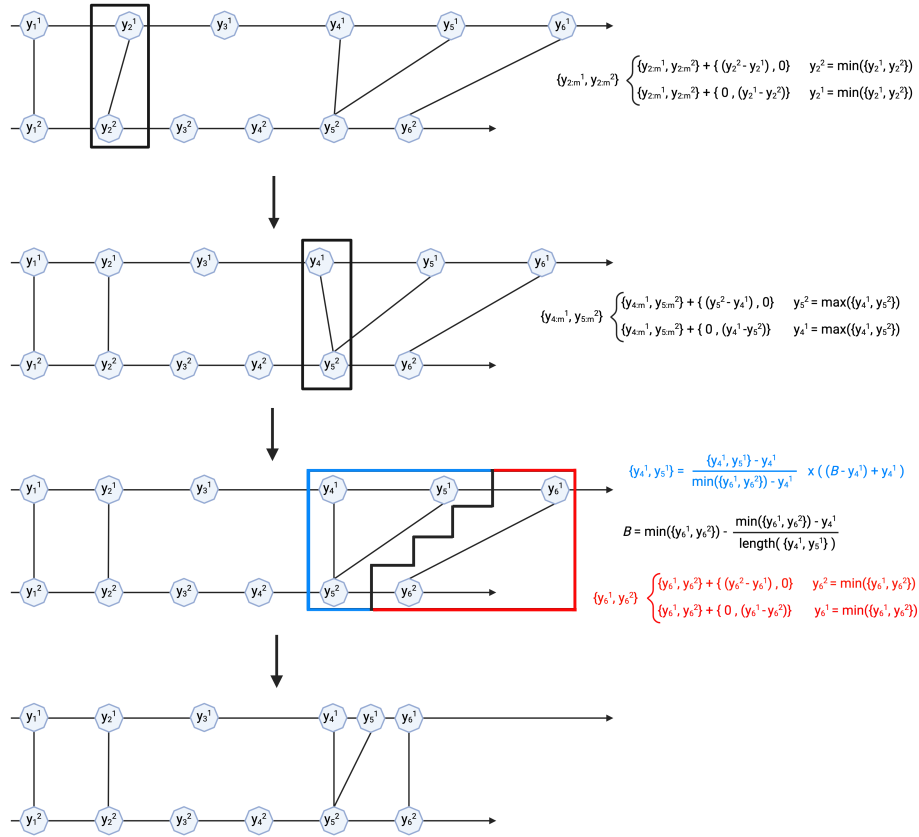
Appendix Figure 2: Graphical explanation of the TrAGEDY process of identifying the optimal start and end points.

TrAGEDY identifies the optimal start and end points by first finding the slice of the score matrix from the start (outlined in blue) and end (outlined in purple) which contains the lowest dissimilarity score match (box coloured orange). TrAGEDY then finds the mean/medians of the score differences between adjacent matches on the start and end slice (A). The means/medians are then used as threshold for considering which matches are the optimal start and end points, with matches whose dissimilarity score is less than the threshold and before the current start match and after the current match (green box) are considered as possible start/end matches (B). For each of the possible paths through the data TrAGEDY performs the following steps. First, it removes any matches (shown in red) that fall outwith the select start and end points and then performs DTW on the cut score matrix. The matches included in the DTW path are then bootstrapped and the mean of the bootstrapped paths is calculated (C).



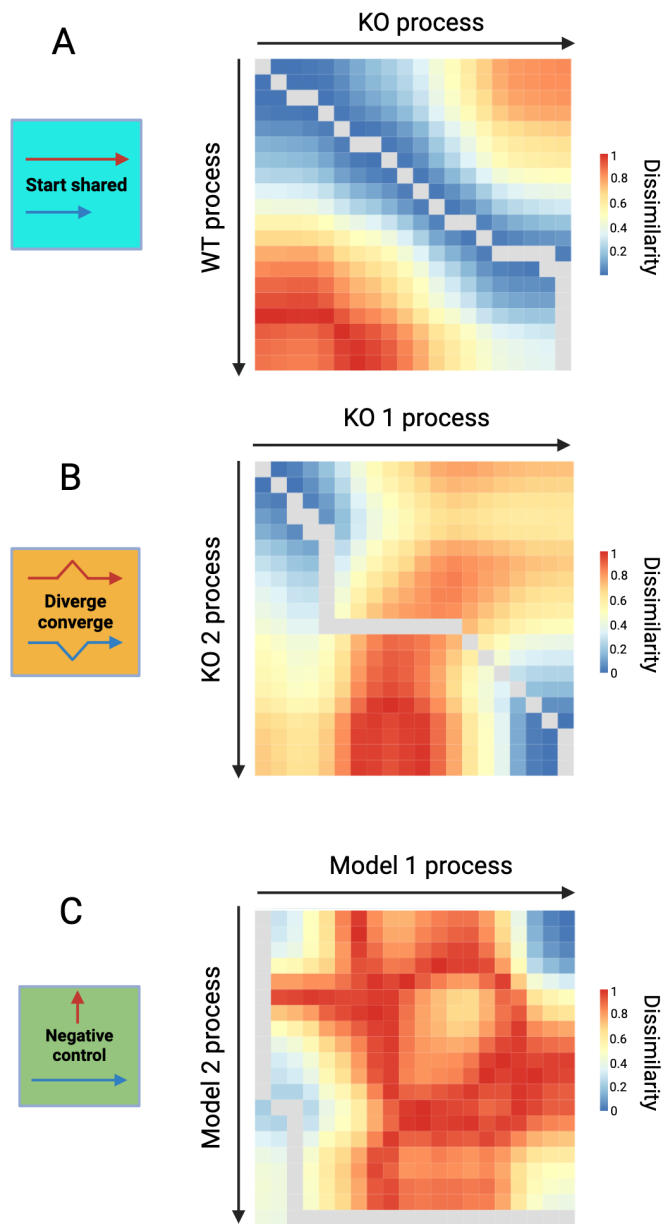
Appendix Figure 3: Graphical explanation of the TrAGEDy process of cutting highly dissimilar matches.

To remove matches from the path which are not optimal TrAGEDy goes through the following process. An unmatched threshold and matched threshold are created by taking the mean/median of scores of the unmatched and matched points respectively (A). Simply, if the absolute difference between a matches score is closer to the unmatched threshold than the matched threshold it is cut and if the opposite is true, it is kept (B).



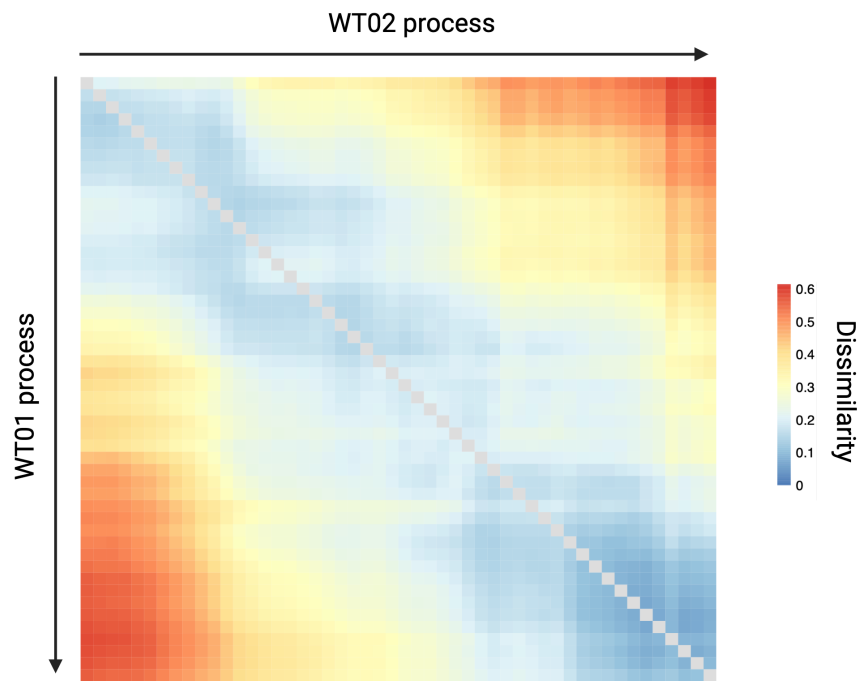
Appendix Figure 4: Graphical explanation and example of the alignment of interpolated point pseudotime values, from the DTW calculated path.

There are two types of movement that can occur when aligning pseudotime values, an individual match (A, B & C) or a multi-match (D). For the individual match, the pseudotime value of the interpolated point with the highest value (and all the subsequent interpolated points) are brought down by the difference between the lower pseudotime value and the higher pseudotime value. For multi-match points, the match that occurs after it is adjusted then the first match in the multi-match. All the interpolated points that are matched to the same interpolated point are then scaled between pseudotime of the initial match in the multi-match and the pseudotime of the next match modified by the difference between the two, normalised by the number of interpolated points to be scaled. This then gives us aligned pseudotime values for all the interpolated points (E).



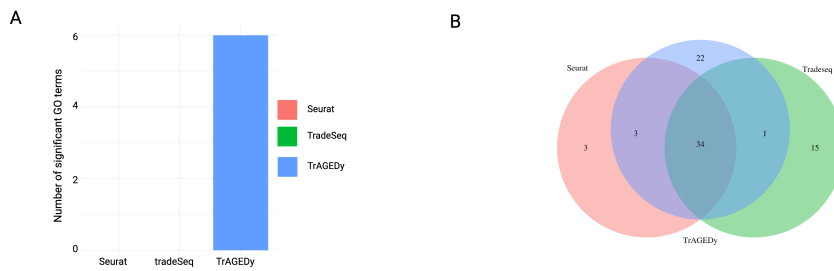
Appendix Figure 5: cellAlign alignments of simulated datasets using Pearson correlation dissimilarity score.

Heatmaps showing the alignment of the two datasets across the three Dyngen simulated alignment topologies. Each box of the heatmap shows the transcriptional dissimilarity (as assessed by Pearson correlation) of interpolated points of the two datasets with blue meaning low dissimilarity and red meaning high dissimilarity. The grey line denotes the optimal alignment of the interpolated points.



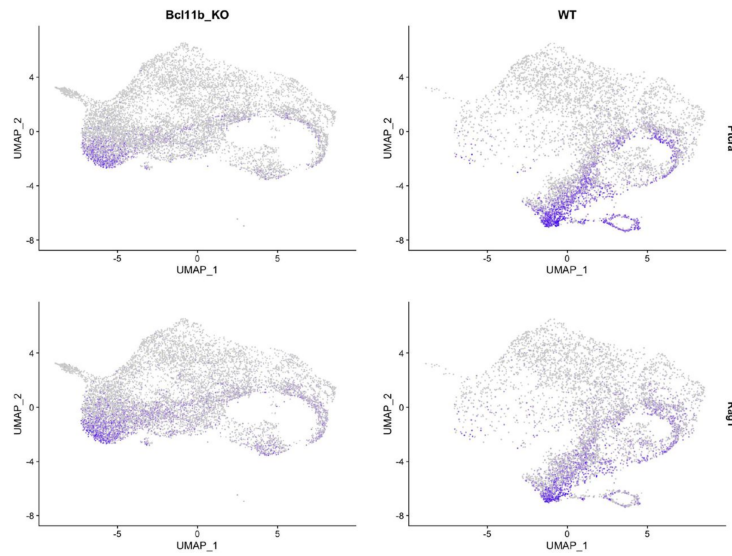
Appendix Figure 6: TrAGEDy alignment of biological replicates of WT slender to stumpy *T. brucei* transition.

TrAGEDy alignment of the WT01 and WT02 slender to stumpy *T. brucei* development trajectories. Dissimilarity in gene expression of interpolated points was calculated using Spearman correlation with blue meaning low dissimilarity and red meaning high dissimilarity.



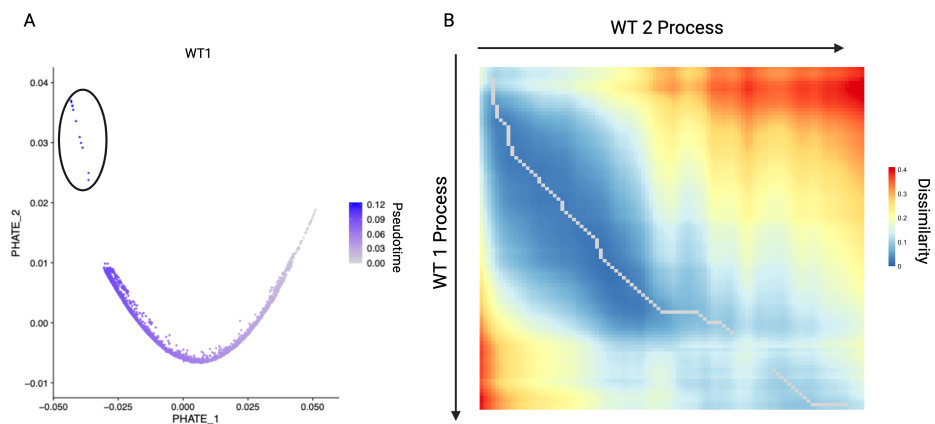
Appendix Figure 7: Further GO term analysis of WT vs *ZC3H20* KO *T. brucei* bloodstream form development.

Barplot showing the number of significant Gene Ontology (GO) terms (Benjamini-Hochberg adjusted p-value < 0.01) returned when GO enrichment analysis was carried out using clusterProfiler on the unique DE genes returned by TrAGEDy, TradeSeq and Seurat (A). Venn diagram showing the intersections of significant GO terms (Benjamini-Hochberg adjusted p-value < 0.01) found by the three methods using clusterProfiler on all the DE genes returned by TrAGEDy, TradeSeq and Seurat (B).



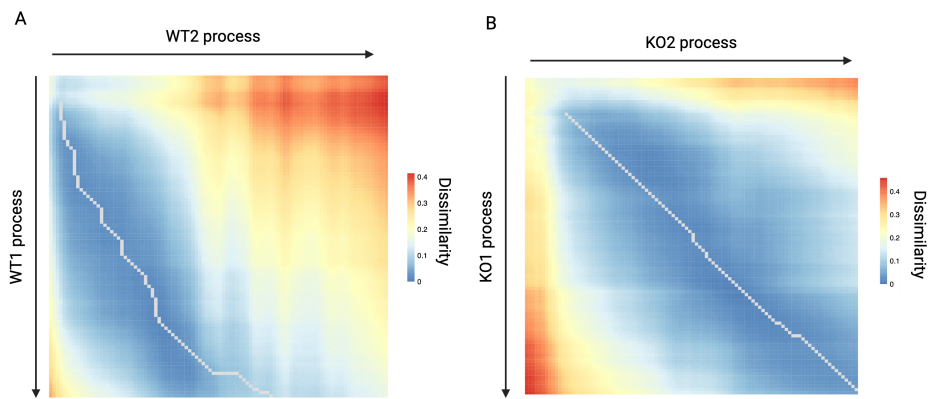
Appendix Figure 8: Expression of markers associated with DN3 stage of T cell development in *Bcl11b* KO and WT T cells.

UMAP showing the normalised gene expression levels of *Ptcra* (pre-T cell receptor α chain) and *Rag1* (recombination activating gene 1) in T cells under WT (right column) and *Bcl11b* KO (left column) conditions.



Appendix Figure 9: Impact of outlier cells on TrAGEDy alignments.

Plot showing the PHATE embeddings of the WT1 T cell dataset with each cell coloured by its cell type annotation. Circle identifies cells which are separate from the main body of the trajectory (A). TrAGEDy alignment of the WT1 and WT2 T cell datasets when the cells circled in A are kept in the trajectory. Dissimilarity in gene expression of interpolated points was calculated using Spearman correlation with blue meaning low dissimilarity and red meaning high dissimilarity.



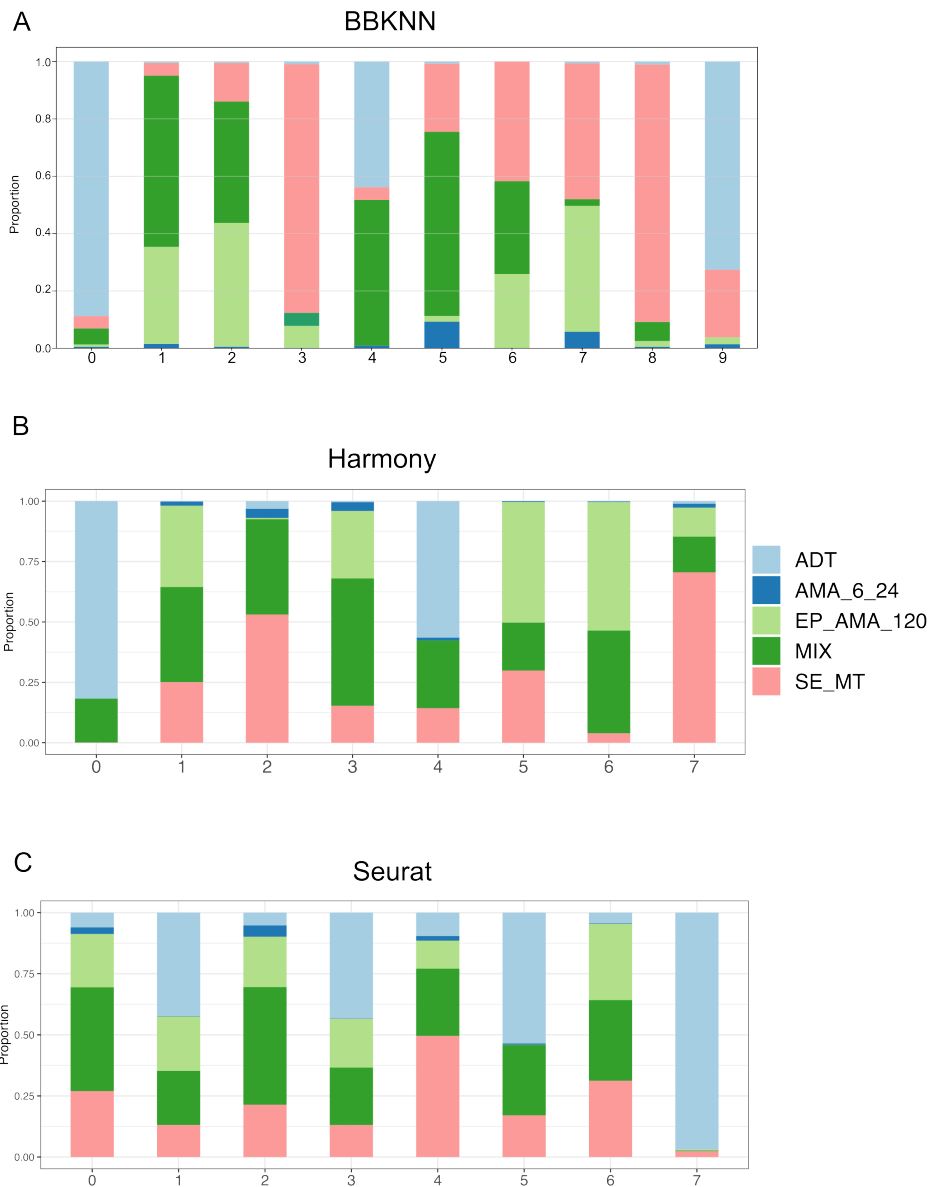
Appendix Figure 10: TrAGEDy alignment of WT and *Bcl11b* KO T cell sequencing run trajectories.

TrAGEDy alignment of the WT1 and WT2 T cell development trajectories (A) and the *Bcl11b* KO1 and *Bcl11b* KO2 T cell development trajectories (B). Dissimilarity in gene expression of interpolated points was calculated using Spearman correlation with blue meaning low dissimilarity and red meaning high dissimilarity.



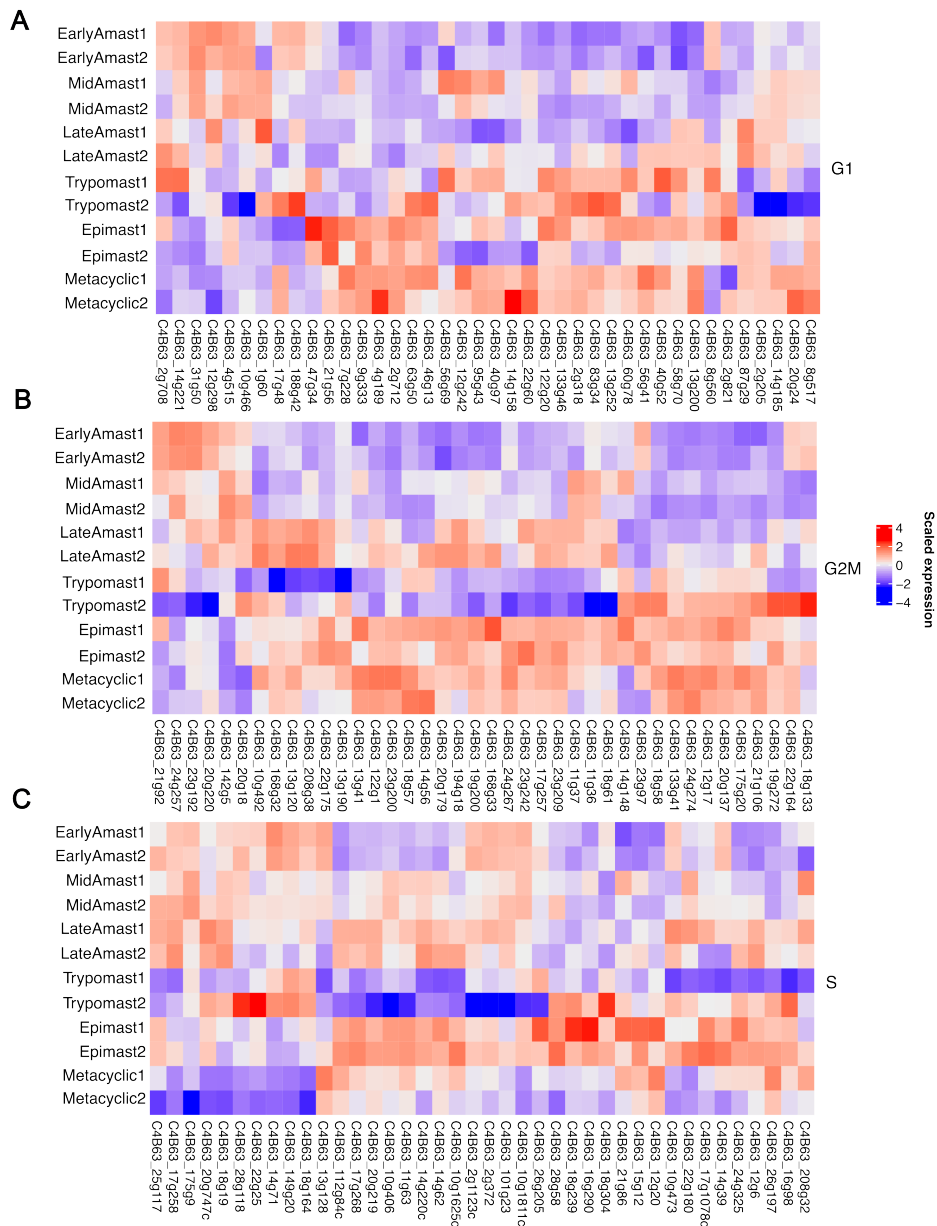
Appendix Figure 11: Further GO term analysis of WT vs *Bcl11b* KO T cell development.

Barplot showing the number of significant Gene Ontology (GO) terms (Benjamini-Hochberg adjusted p-value < 0.01) returned when GO enrichment analysis was carried out using clusterProfiler on the unique DE genes returned by TrAGEDy, TradeSeq and Seurat (A). Venn diagram showing the intersections of significant GO terms (Benjamini-Hochberg adjusted p-value < 0.01) found by the three methods using clusterProfiler on all the DE genes returned by TrAGEDy, TradeSeq and Seurat (B).



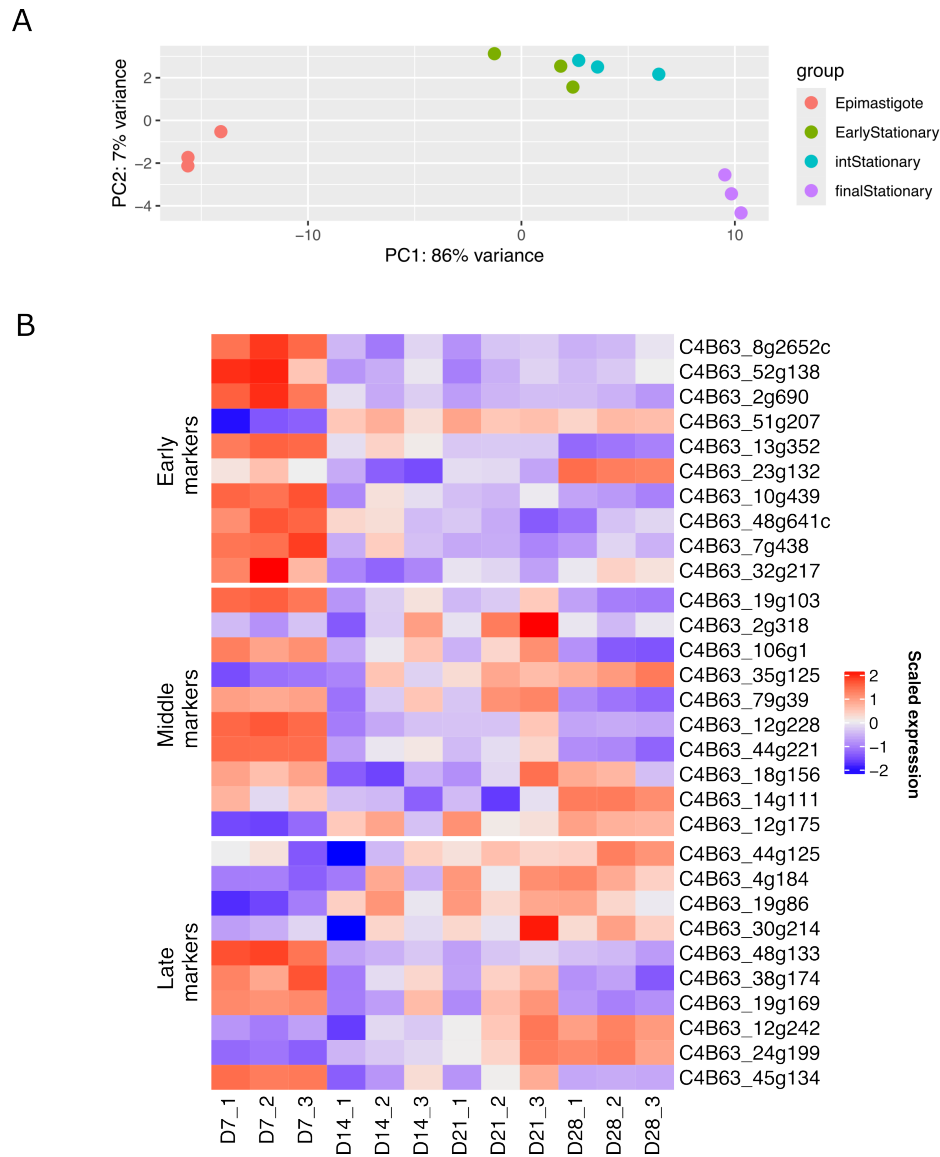
Appendix Figure 12: *T. cruzi* atlas integration method testing

Proportion plots showing the distribution of cells from the five main sampling strategies (the MIX1 and MIX2 and SE/MT 1 and SE/MT 2 replicates have been collapsed together) across the clusters generated from the BBKNN (A), Harmony (B) and Seurat V5 CCA (C) integration methods. The sampling strategy legend for the entire figure can be seen the right of panel B.



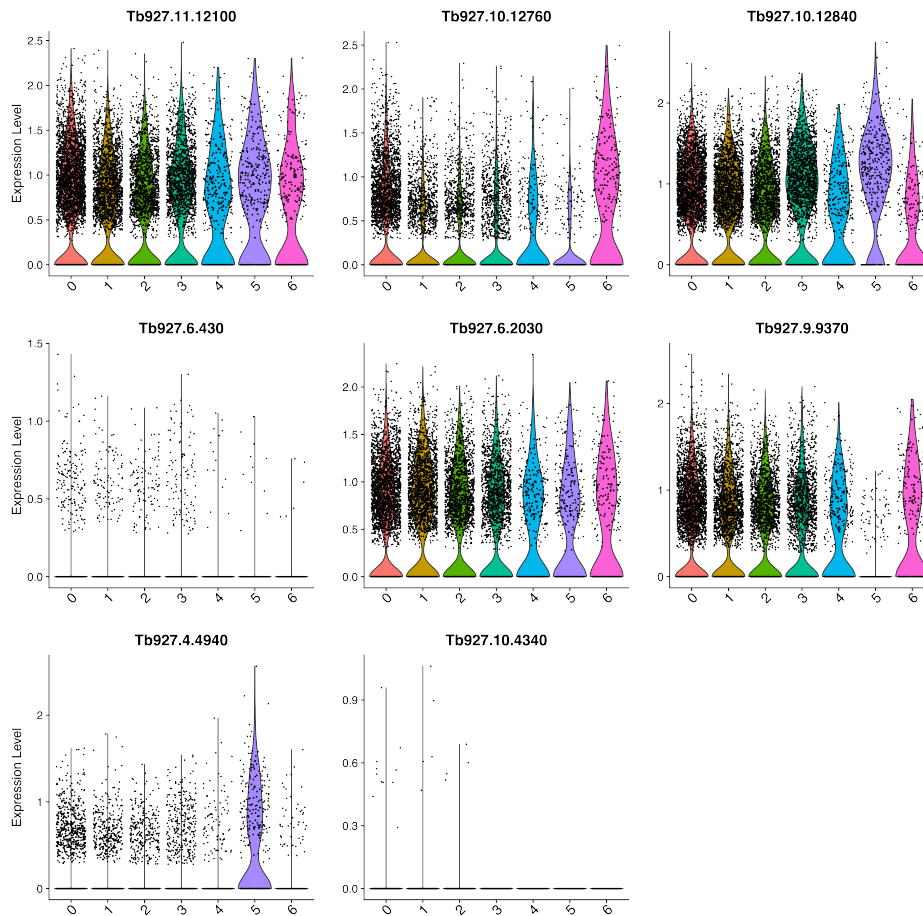
Appendix Figure 13: Expression of cell cycle phase marker genes in the bulk RNA-seq datasets

Heatmaps showing the the scaled normalised expression values across the *T. cruzi* bulk RNA-seq datasets for 50 randomly selected marker genes of the G1 (A), G2M (B) and S (C) cell cycle phases. The genes were taken from Chávez and colleagues (Chávez, Urbaniak, et al. 2021)



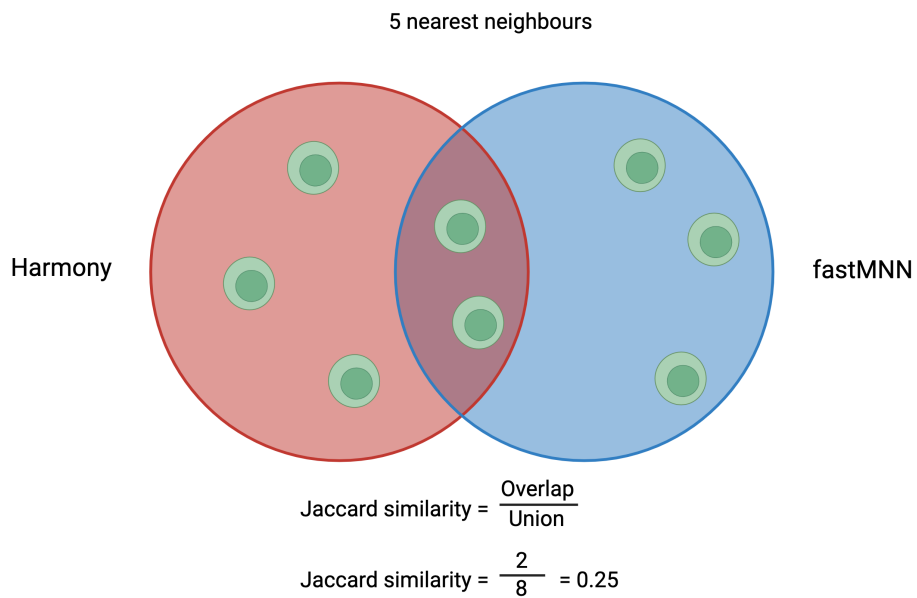
Appendix Figure 14: **scRNA-seq metacyclogenesis marker gene expression in bulk RNA-seq samples of metacyclogenesis**

PCA plot showing the bulk RNA-seq datasets coloured by the four different time points of metacyclogenesis sampled by Smircich and colleagues (Smircich, Perez-Diaz, et al. 2023) (A). Heatmap showing the scaled normalised expression values for the metacyclogenesis bulk RNA-seq datasets of the top 10 (in terms of Log_2FC) genes found to be significantly associated with metacyclogenesis from the scRNA-seq analysis (B).



Appendix Figure 15: **Expression of mutated *RBP6* overexpression cell marker genes**

Violin plots of the scRNA-seq *RBP6* overexpression dataset, showing the normalized expression values of the eight transcripts that were identified by Shi and colleagues (H. Shi, K. Butler, and Tschudi 2018) as upregulated over the *RBP6* overexpression process in the mutated *RBP6* form but not the WT *RBP6* form.



Appendix Figure 16: **Graphical explanation of Jaccard similarity**

Jaccard similarity is calculated by finding the overlap and union of two lists and dividing them. In the context of this chapter, the lists are the nearest neighbours of each cell in the integrated space of two integration methods. In the example in this figure, the 5 nearest neighbours of a cell in the Harmony integrated space and the fastMNN integrated space.

25 References

References

- Acosta-Serrano, A., E. Vassella, M. Liniger, C.K. Renggli, R. Brun, I. Roditi, and P.T. Englund (2001). “The surface coat of procyclic *Trypanosoma brucei*: Programmed expression and proteolytic cleavage of procyclin in the tsetse fly”. In: *Proceedings of the National Academy of Sciences* 98.4, pp. 1513–1518.
- Adroher, F., A. Osuna, and J. A. Lupiañez (1988). “Differential Energetic Metabolism during *Trypanosoma cruzi* Differentiation”. In: *Archives of Biochemistry and Biophysics* 267.1, pp. 252–261.
- Ahlmann-Eltze, C. and W. Huber (2023). “Comparison of transformations for single-cell RNA-seq data”. In: *Nat Methods* 20.5, pp. 665–672. URL: <https://www.ncbi.nlm.nih.gov/pubmed/37037999>.
- Aksoy, E., A. Vigneron, X. Bing, X. Zhao, M. O’Neill, Y. N. Wu, J. D. Bangs, B. L. Weiss, and S. Aksoy (2016). “Mammalian African trypanosome VSG coat enhances tsetse’s vector competence”. In: *Proc Natl Acad Sci U S A* 113.25, pp. 6961–6. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27185908>.
- Aksoy, S. (2019). “Tsetse peritrophic matrix influences for trypanosome transmission”. In: *J Insect Physiol* 118, p. 103919. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31425686>.
- Alivernini, S. et al. (2020). “Distinct synovial tissue macrophage subsets regulate inflammation and remission in rheumatoid arthritis”. In: *Nat Med* 26.8, pp. 1295–1306. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32601335>.
- Alpert, A., L. S. Moore, T. Dubovik, and S. S. Shen-Orr (2018). “Alignment of single-cell trajectories to compare cellular expression dynamics”. In: *Nat Methods* 15.4, pp. 267–270. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29529018>.
- Alvarez, V. E., G. Kosec, C. Sant’Anna, V. Turk, J. J. Cazzulo, and B. Turk (2008). “Autophagy is involved in nutritional stress response and differentiation in *Trypanosoma cruzi*”. In: *J Biol Chem* 283.6, pp. 3454–3464. URL: <https://www.ncbi.nlm.nih.gov/pubmed/18039653>.
- Alvarez-Jarreta, J. et al. (2024). “VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center in 2023”. In: *Nucleic Acids Res* 52.D1, pp. D808–D816. URL: <https://www.ncbi.nlm.nih.gov/pubmed/37953350>.
- Amezquita, R. A., A. T. L. Lun, E. Becht, V. J. Carey, L. N. Carpp, L. Geistlinger, F. Marini, K. Rue-Albrecht, D. Risso, C. Sonesson, L. Waldron, H. Pages, M. L. Smith, W. Huber, M. Morgan, R. Gottardo, and S. C. Hicks (2020). “Orchestrating single-cell analysis with Bioconductor”. In: *Nat Methods* 17.2, pp. 137–145. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31792435>.
- Amodeo, S., I. Bregy, and T. Ochsenreiter (2023). “Mitochondrial genome maintenance—the kinetoplast story”. In: *FEMS Microbiol Rev* 47.6. URL: <https://www.ncbi.nlm.nih.gov/pubmed/36449697>.
- Amos, B. et al. (2022). “VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center”. In: *Nucleic Acids Res* 50.D1, pp. D898–D911. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34718728>.
- Anders, S. and W. Huber (2010). “Differential expression analysis for sequence count data”. In: *Genome Biology* 11.106, pp. 1–12.
- Andrews, N.W. and M.B. Whitlow (1989). “Secretion by *Trypanosoma cruzi* of a hemolysin active at low pH”. In: *Molecular and Biochemical Parasitology* 33, pp. 249–256.
- Aran, D., A. P. Looney, L. Liu, E. Wu, V. Fong, A. Hsu, S. Chak, R. P. Naikawadi, P. J. Wolters, A. R. Abate, A. J. Butte, and M. Bhattacharya (2019). “Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage”. In: *Nat Immunol* 20.2, pp. 163–172. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30643263>.

- Archer, S. K., V. D. Luu, R. A. de Queiroz, S. Brems, and C. Clayton (2009). “Trypanosoma brucei PUF9 regulates mRNAs for proteins involved in replicative processes over the cell cycle”. In: *PLoS Pathog* 5.8, e1000565. URL: <https://www.ncbi.nlm.nih.gov/pubmed/19714224>.
- Baltz, T., D. Baltz, C.H. Giroud, and J. Crockett (1985). “Cultivation in a semi-defined medium of animal infective forms of Trypanosoma brucei, T. equiperdum, T. evansi, T. rhodesiense and T. gambiense”. In: *The EMBO Journal* 4.5, pp. 1273–1277.
- Barisón, M. J., L. N. Rapado, E. F. Merino, E. M. Furusho Pral, B. S. Mantilla, L. Marchese, C. Nowicki, A. M. Silber, and M. B. Cassera (2017). “Metabolomic profiling reveals a finely tuned, starvation-induced metabolic switch in Trypanosoma cruzi epimastigotes”. In: *J Biol Chem* 292.21, pp. 8964–8977. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28356355>.
- Barry, J.D. and R. McCulloch (2001). “Antigenic variation in trypanosomes: Enhanced phenotypic variation in a eukaryotic parasite”. In: *Advances in Parasitology* 49, pp. 1–70.
- Batista, M. F., C. A. Najera, I. Meneghelli, and D. Bahia (2020). “The Parasitic Intracellular Lifestyle of Trypanosomatids: Parasitophorous Vacuole Development and Survival”. In: *Front Cell Dev Biol* 8, p. 396. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32587854>.
- Belew, A. T., C. Junqueira, G. F. Rodrigues-Luiz, B. M. Valente, A. E. R. Oliveira, R. B. Polidoro, L. W. Zuccherato, D. C. Bartholomeu, S. Schenkman, R. T. Gazzinelli, B. A. Burleigh, N. M. El-Sayed, and S. M. R. Teixeira (2017). “Comparative transcriptome profiling of virulent and non-virulent Trypanosoma cruzi underlines the role of surface proteins during infection”. In: *PLoS Pathog* 13.12, e1006767. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29240831>.
- Berna, L., M. Rodriguez, M. L. Chiribao, A. Parodi-Talice, S. Pita, G. Rijo, F. Alvarez-Valin, and C. Robello (2018). “Expanding an expanded genome: long-read sequencing of Trypanosoma cruzi”. In: *Microb Genom* 4.5. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29708484>.
- Berriman, M. et al. (2005). “The Genome of the African Trypanosome Trypanosoma brucei”. In: *Science* 309.
- Betz, B. C., K. L. Jordan-Williams, C. Wang, S. G. Kang, J. Liao, M. R. Logan, C. H. Kim, and E. J. Taparowsky (2010). “Batf coordinates multiple aspects of B and T cell function required for normal antibody responses”. In: *J Exp Med* 207.5, pp. 933–42. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20421391>.
- Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre (2008). “Fast unfolding of communities in large networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10.
- Booeshaghi, A. Sina, Ingileif B. Hallgrímsdóttir, Ángel Gálvez-Merchán, and Lior Pachter (2022). “Depth normalization for single-cell genomics count data”. In: *bioRxiv*.
- Brener, Z. and E Chiari (1963). “Observations on the chronic phase of experimental Chagas’ disease in mice”. In: *Revista do Instituto de Medicina Tropical de São Paulo* 5, pp. 128–132.
- Briggs, E. M., C. A. Marques, G. R. Oldrieve, J. Hu, T. D. Otto, and K. R. Matthews (2023). “Profiling the bloodstream form and procyclic form Trypanosoma brucei cell cycle using single-cell transcriptomics”. In: *Elife* 12. URL: <https://www.ncbi.nlm.nih.gov/pubmed/37166108>.
- Briggs, E. M., F. Rojas, R. McCulloch, K. R. Matthews, and T. D. Otto (2021). “Single-cell transcriptomic analysis of bloodstream Trypanosoma brucei reconstructs cell cycle progression and developmental quorum sensing”. In: *Nat Commun* 12.1, p. 5268. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34489460>.
- Burkard, G.S., P. Jutzi, and I. Roditi (2011). “Genome-wide RNAi screens in bloodstream form trypanosomes identify drug transporters”. In: *Mol Biochem Parasitol* 175.1, pp. 91–4. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20851719>.
- Butler, A., P. Hoffman, P. Smibert, E. Papalexi, and R. Satija (2018). “Integrating single-cell transcriptomic data across different conditions, technologies, and species”. In: *Nat Biotechnol* 36.5, pp. 411–420. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29608179>.

- Cannoodt, R., W. Saelens, L. Deconinck, and Y. Saeys (2021). “Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells”. In: *Nat Commun* 12.1, p. 3942. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34168133>.
- Cano-Gamez, E., B. Soskic, T. I. Roumeliotis, E. So, D. J. Smyth, M. Baldrighi, D. Wille, N. Nakic, J. Esparza-Gordillo, C. G. C. Larminie, P. G. Bronson, D. F. Tough, W. C. Rowan, J. S. Choudhary, and G. Trynka (2020). “Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4(+) T cells to cytokines”. In: *Nat Commun* 11.1, p. 1801. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32286271>.
- Cao, J., M. Spielmann, X. Qiu, X. Huang, D. M. Ibrahim, A. J. Hill, F. Zhang, S. Mundlos, L. Christiansen, F. J. Steemers, C. Trapnell, and J. Shendure (2019). “The single-cell transcriptional landscape of mammalian organogenesis”. In: *Nature* 566.7745, pp. 496–502. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30787437>.
- Caradonna, K. L., J. C. Engel, D. Jacobi, C. H. Lee, and B. A. Burleigh (2013). “Host metabolism regulates intracellular growth of *Trypanosoma cruzi*”. In: *Cell Host Microbe* 13.1, pp. 108–17. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23332160>.
- Caspi, R., T. Altman, R. Billington, K. Dreher, H. Foerster, C. A. Fulcher, T. A. Holland, I. M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, D. Weerasinghe, P. Zhang, and P. D. Karp (2014). “The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases”. In: *Nucleic Acids Res* 42.Database issue, pp. D459–71. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24225315>.
- Cattell, R. B. (1966). “The Scree Test For The Number Of Factors”. In: *Multivariate Behav Res* 1.2, pp. 245–76. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26828106>.
- Cayla, M., L. McDonald, P. MacGregor, and K. Matthews (2020). “An atypical DYRK kinase connects quorum-sensing with posttranscriptional gene regulation in *Trypanosoma brucei*”. In: *Elife* 9. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32213288>.
- Chamond, N., M. Goytia, N. Coatnoan, J. C. Barale, A. Cosson, W. M. Degrave, and P. Minoprio (2005). “*Trypanosoma cruzi* proline racemases are involved in parasite differentiation and infectivity”. In: *Mol Microbiol* 58.1, pp. 46–60. URL: <https://www.ncbi.nlm.nih.gov/pubmed/16164548>.
- Chari, T. and L. Pachter (2023). “The specious art of single-cell genomics”. In: *PLoS Comput Biol* 19.8, e1011288. URL: <https://www.ncbi.nlm.nih.gov/pubmed/37590228>.
- Chávez, S., G. Eastman, P. Smircich, L. L. Becco, C. Oliveira-Rizzo, R. Fort, M. Potenza, B. Garat, J. R. Sotelo-Silveira, and M. A. Duhagon (2017). “Transcriptome-wide analysis of the *Trypanosoma cruzi* proliferative cycle identifies the periodically expressed mRNAs and their multiple levels of control”. In: *PLoS One* 12.11, e0188441. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29182646>.
- Chávez, S., M. D. Urbaniak, C. Benz, P. Smircich, B. Garat, J. R. Sotelo-Silveira, and M. A. Duhagon (2021). “Extensive Translational Regulation through the Proliferative Transition of *Trypanosoma cruzi* Revealed by Multi-Omics”. In: *mSphere* 6.5, e0036621. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34468164>.
- Chen, H., L. Albergante, J. Y. Hsu, C. A. Lareau, G. Lo Bosco, J. Guan, S. Zhou, A. N. Gorban, D. E. Bauer, M. J. Aryee, D. M. Langenau, A. Zinovyev, J. D. Buenrostro, G. C. Yuan, and L. Pinello (2019). “Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM”. In: *Nat Commun* 10.1, p. 1903. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31015418>.
- Cheng, Y., X. Fan, J. Zhang, and Y. Li (2023). “A scalable sparse neural network framework for rare cell type annotation of single-cell transcriptome data”. In: *Commun Biol* 6.1, p. 545. URL: <https://www.ncbi.nlm.nih.gov/pubmed/37210444>.
- Christiano, R., N. G. Kolev, H. Shi, E. Ullu, T. C. Walther, and C. Tschudi (2017). “The proteome and transcriptome of the infectious metacyclic form of *Trypanosoma brucei* define quiescent cells primed

- for mammalian invasion”. In: *Mol Microbiol* 106.1, pp. 74–92. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28742275>.
- Cinar, B., P. K. Fang, M. Lutchman, D. Di Vizio, R. M. Adam, N. Pavlova, M. A. Rubin, P. C. Yelick, and M. R. Freeman (2007). “The pro-apoptotic kinase Mst1 and its caspase cleavage products are direct inhibitors of Akt1”. In: *EMBO J* 26.21, pp. 4523–34. URL: <https://www.ncbi.nlm.nih.gov/pubmed/17932490>.
- Clayton, C. (2019). “Regulation of gene expression in trypanosomatids: living with polycistronic transcription”. In: *Open Biol* 9.6, p. 190072. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31164043>.
- Contreras, V.T., J.M. Salles, N. Thomas, C.M. Morel, and S. Goldenberg (1985). “In vitro differentiation of trypanosoma cruzi under chemically defined conditions”. In: *Molecular and Biochemical Parasitology* 16, pp. 315–327.
- Coughlin, B. C., S. M. Teixeira, L. V. Kirchhoff, and J. E. Donelson (2000). “Amastin mRNA abundance in *Trypanosoma cruzi* is controlled by a 3′-untranslated region position-dependent cis-element and an untranslated region-binding protein”. In: *J Biol Chem* 275.16, pp. 12051–60. URL: <https://www.ncbi.nlm.nih.gov/pubmed/10766837>.
- Cross, G.A.M. and J.C. Manning (1973). “Cultivation of *Trypanosoma brucei* spp. in semi-defined and defined media”. In: *Parasitology* 67.3, pp. 315–331.
- Cunningham, I. (1977). “New culture medium for maintenance of tsetse tissues and growth of trypanosomatids”. In: *J Protozool* 24.2, pp. 325–9. URL: <https://www.ncbi.nlm.nih.gov/pubmed/881656>.
- Czichos, J., C. Nonnengaesser, and P. Overath (1986). “*Trypanosoma brucei*: cis-aconitate and temperature reduction as triggers of synchronous transformation of bloodstream to procyclic trypomastigotes in vitro”. In: *Experimental Parasitology* 62.2, pp. 283–91.
- Das, A., H. Li, T. Liu, and V. Bellofatto (2006). “Biochemical characterization of *Trypanosoma brucei* RNA polymerase II”. In: *Mol Biochem Parasitol* 150.2, pp. 201–10. URL: <https://www.ncbi.nlm.nih.gov/pubmed/16962183>.
- De Donno, C., S. Hediye-Zadeh, A. A. Moinfar, M. Wagenstetter, L. Zappia, M. Lotfollahi, and F. J. Theis (2023). “Population-level integration of single-cell datasets enables multi-scale analysis across samples”. In: *Nat Methods* 20.11, pp. 1683–1692. URL: <https://www.ncbi.nlm.nih.gov/pubmed/37813989>.
- De Pablos, L. M., G. G. Gonzalez, J. Solano Parada, V. Seco Hidalgo, I. M. Diaz Lozano, M. M. Gomez Samblas, T. Cruz Bustos, and A. Osuna (2011). “Differential expression and characterization of a member of the mucin-associated surface protein family secreted by *Trypanosoma cruzi*”. In: *Infect Immun* 79.10, pp. 3993–4001. URL: <https://www.ncbi.nlm.nih.gov/pubmed/21788387>.
- Dean, S., R. Marchetti, K. Kirk, and K. R. Matthews (2009). “A surface transporter family conveys the trypanosome differentiation signal”. In: *Nature* 459.7244, pp. 213–7. URL: <https://www.ncbi.nlm.nih.gov/pubmed/19444208>.
- Devaux, S., L. Lecordier, P. Uzureau, D. Walgraffe, J. F. Dierick, P. Poelvoorde, E. Pays, and L. Vanhamme (2006). “Characterization of RNA polymerase II subunits of *Trypanosoma brucei*”. In: *Mol Biochem Parasitol* 148.1, pp. 60–8. URL: <https://www.ncbi.nlm.nih.gov/pubmed/16621069>.
- Diaz Soria, C. L., T. Attenborough, Z. Lu, S. Fontenla, J. Graham, C. Hall, S. Thompson, T. G. R. Andrews, K. A. Rawlinson, M. Berriman, and G. Rinaldi (2024). “Single-cell transcriptomics of the human parasite *Schistosoma mansoni* first intra-molluscan stage reveals tentative tegumental and stem-cell regulators”. In: *Sci Rep* 14.1, p. 5974. URL: <https://www.ncbi.nlm.nih.gov/pubmed/38472267>.
- Diaz Soria, C. L., J. Lee, T. Chong, A. Coghlan, A. Tracey, M. D. Young, T. Andrews, C. Hall, B. L. Ng, K. Rawlinson, S. R. Doyle, S. Leonard, Z. Lu, H. M. Bennett, G. Rinaldi, P. A. Newmark, and M. Berriman (2020). “Single-cell atlas of the first intra-mammalian developmental stage of the human parasite *Schistosoma mansoni*”. In: *Nat Commun* 11.1, p. 6411. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33339816>.

- Dijk, E. L. van, Y. Jaszczyszyn, and C. Thermes (2014). “Library preparation methods for next-generation sequencing: tone down the bias”. In: *Exp Cell Res* 322.1, pp. 12–20. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24440557>.
- Dixit, A., O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman, N. Friedman, and A. Regev (2016). “Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens”. In: *Cell* 167.7, 1853–1866 e17. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27984732>.
- Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras (2013). “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1, pp. 15–21. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23104886>.
- Dolezelova, E., M. Kunzova, M. Dejung, M. Levin, B. Panicucci, C. Regnault, C. J. Janzen, M. P. Barrett, F. Butter, and A. Zikova (2020). “Cell-based and multi-omics profiling reveals dynamic metabolic repurposing of mitochondria to drive developmental progression of *Trypanosoma brucei*”. In: *PLoS Biol* 18.6, e3000741. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32520929>.
- Dooley, N. L., T. G. Chabikwa, Z. Pava, J. R. Loughland, J. Hamelink, K. Berry, D. Andrew, M. S. F. Soon, A. SheelaNair, K. A. Piera, T. William, B. E. Barber, M. J. Grigg, C. R. Engwerda, J. A. Lopez, N. M. Anstey, and M. J. Boyle (2023). “Single cell transcriptomics shows that malaria promotes unique regulatory responses across multiple immune cell subsets”. In: *Nat Commun* 14.1, p. 7387. URL: <https://www.ncbi.nlm.nih.gov/pubmed/37968278>.
- Dumoulin, P. C., J. Vollrath, S. S. Tomko, J. X. Wang, and B. Burleigh (2020). “Glutamine metabolism modulates azole susceptibility in *Trypanosoma cruzi* amastigotes”. In: *Elife* 9. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33258448>.
- Durieux, P.O., P. Schültz, R. Brun, and P. Köhler (1991). “Alterations in Krebs cycle enzyme activities and carbohydrate catabolism in two strains of T13’panosomabruceiduring in vitro differentiation of their bloodstream to procyclic stages”. In: *Molecular and Biochemical Parasitology* 45, pp. 19–28.
- Engstler, M. and M. Boshart (2004). “Cold shock and regulation of surface protein trafficking convey sensitization to inducers of stage differentiation in *Trypanosoma brucei*”. In: *Genes Dev* 18.22, pp. 2798–811. URL: <https://www.ncbi.nlm.nih.gov/pubmed/15545633>.
- Erben, E., K. Leiss, B. Liu, D. I. Gil, C. Helbig, and C. Clayton (2021). “Insights into the functions and RNA binding of *Trypanosoma brucei* ZC3H22, RBP9 and DRBD7”. In: *Parasitology* 148.10, pp. 1186–1195. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33536101>.
- Ester, M., H. Kriegel, J. Sander, and X. Xu (1996). “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: *Association for the Advancement of Artificial Intelligence*, pp. 226–231.
- Evans, D. A. and R. C. Brown (1972). “The utilization of glucose and proline by culture forms of *Trypanosoma brucei*”. In: *J Protozool* 19.4, pp. 686–90. URL: <https://www.ncbi.nlm.nih.gov/pubmed/4641906>.
- Fadda, A., M. Ryten, D. Droll, F. Rojas, V. Farber, J. R. Haanstra, C. Merce, B. M. Bakker, K. Matthews, and C. Clayton (2014). “Transcriptome-wide analysis of trypanosome mRNA decay reveals complex degradation kinetics and suggests a role for co-transcriptional degradation in determining mRNA levels”. In: *Mol Microbiol* 94.2, pp. 307–26. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25145465>.
- Fernandez-Moya, S. M. and A. M. Estevez (2010). “Posttranscriptional control and the role of RNA-binding proteins in gene regulation in trypanosomatid protozoan parasites”. In: *Wiley Interdiscip Rev RNA* 1.1, pp. 34–46. URL: <https://www.ncbi.nlm.nih.gov/pubmed/21956905>.
- Ferreira, E. R., A. Bonfim-Melo, B. A. Burleigh, J. A. Costales, K. M. Tyler, and R. A. Mortara (2021). “Parasite-Mediated Remodeling of the Host Microfilament Cytoskeleton Enables Rapid Egress of

- Trypanosoma cruzi following Membrane Rupture”. In: *mBio* 12.3, e0098821. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34154418>.
- Fetene, E., S. Leta, F. Regassa, and P. Buscher (2021). “Global distribution, host range and prevalence of Trypanosoma vivax: a systematic review and meta-analysis”. In: *Parasit Vectors* 14.1, p. 80. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33494807>.
- Florini, F., A. Naguleswaran, W. H. Gharib, F. Bringaud, and I. Roditi (2019). “Unexpected diversity in eukaryotic transcription revealed by the retrotransposon hotspot family of Trypanosoma brucei”. In: *Nucleic Acids Res* 47.4, pp. 1725–1739. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30544263>.
- Fonseca, L. M. da, K. M. da Costa, V. S. Chaves, C. G. Freire-de-Lima, A. Morrot, L. Mendonca-Previato, J. O. Previato, and L. Freire-de-Lima (2019). “Theft and Reception of Host Cell’s Sialic Acid: Dynamics of Trypanosoma Cruzi Trans-sialidases and Mucin-Like Molecules on Chagas’ Disease Immunomodulation”. In: *Front Immunol* 10, p. 164. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30787935>.
- Franzen, O., S. Ochaya, E. Sherwood, M. D. Lewis, M. S. Llewellyn, M. A. Miles, and B. Andersson (2011). “Shotgun sequencing analysis of Trypanosoma cruzi I Sylvio X10/1 and comparison with T. cruzi VI CL Brener”. In: *PLoS Negl Trop Dis* 5.3, e984. URL: <https://www.ncbi.nlm.nih.gov/pubmed/21408126>.
- Garcia, E. S., F. A. Genta, P. de Azambuja, and G. A. Schaub (2010). “Interactions between intestinal compounds of triatomines and Trypanosoma cruzi”. In: *Trends Parasitol* 26.10, pp. 499–505. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20801082>.
- Garcia, E.S. and P. Azambuja (1991). “Development and Interactions of Trypanosomacruzi/Within the Insect Vector”. In: *Parasitology Today* 7.9, pp. 240–244.
- Garcia-Huertas, P., Y. Cuesta-Astroz, V. Araque-Ruiz, and N. Cardona-Castro (2023). “Transcriptional changes during metacyclogenesis of a Colombian Trypanosoma cruzi strain”. In: *Parasitol Res* 122.2, pp. 625–634. URL: <https://www.ncbi.nlm.nih.gov/pubmed/36567399>.
- Gazos-Lopes, F., J. L. Martin, P. C. Dumoulin, and B. A. Burleigh (2017). “Host triacylglycerols shape the lipidome of intracellular trypanosomes and modulate their growth”. In: *PLoS Pathog* 13.12, e1006800. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29281741>.
- Genomics, 10X (2017). “8k PBMCs from a Healthy Donor”. In: URL: <https://www.10xgenomics.com/datasets/8-k-pbm-cs-from-a-healthy-donor-2-standard-2-1-0>.
- (2018). *What is the maximum number of cells that can be profiled?* Web Page. URL: <https://kb.10xgenomics.com/hc/en-us/articles/360001378811-What-is-the-maximum-number-of-cells-that-can-be-profiled>.
- Giotti, B., S. H. Chen, M. W. Barnett, T. Regan, T. Ly, S. Wiemann, D. A. Hume, and T. C. Freeman (2019). “Assembly of a parts list of the human mitotic cell cycle machinery”. In: *J Mol Cell Biol* 11.8, pp. 703–718. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30452682>.
- Gomes Passos Silva, D. et al. (2018). “The in vivo and in vitro roles of Trypanosoma cruzi Rad51 in the repair of DNA double strand breaks and oxidative lesions”. In: *PLoS Negl Trop Dis* 12.11, e0006875. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30422982>.
- Goncalves, C. S., A. R. Avila, W. de Souza, M. C. M. Motta, and D. P. Cavalcanti (2018). “Revisiting the Trypanosoma cruzi metacyclogenesis: morphological and ultrastructural analyses during cell differentiation”. In: *Parasit Vectors* 11.1, p. 83. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29409544>.
- Gonzales-Perdomo, M., P. Romero, and S. Goldenberg (1988). “Cyclic AMP and Adenylate Cyclase Activators Stimulate Ttrypanosoma cruzi Differentiation”. In: *Experimental Parasitology* 66, pp. 205–212.
- Gonzalez, M. S., M. S. Souza, E. S. Garcia, N. F. Nogueira, C. B. Mello, G. E. Canepa, S. Bertotti, I. M. Durante, P. Azambuja, and C. A. Buscaglia (2013). “Trypanosoma cruzi TcSMUG L-surface mucins promote development and infectivity in the triatomine vector Rhodnius prolixus”. In: *PLoS Negl Trop Dis* 7.11, e2552. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24244781>.

- Gorin, G., M. Fang, T. Chari, and L. Pachter (2022). “RNA velocity unraveled”. In: *PLoS Comput Biol* 18.9, e1010492. URL: <https://www.ncbi.nlm.nih.gov/pubmed/36094956>.
- Grinsven, K. W. van, J. Van Den Abbeele, P. Van den Bossche, J. J. van Hellemond, and A. G. Tielens (2009). “Adaptations in the glucose metabolism of procyclic *Trypanosoma brucei* isolates from tsetse flies and during differentiation of bloodstream forms”. In: *Eukaryot Cell* 8.8, pp. 1307–11. URL: <https://www.ncbi.nlm.nih.gov/pubmed/19542311>.
- Grisard, E. C., S. M. Teixeira, L. G. de Almeida, P. H. Stoco, A. L. Gerber, C. Talavera-Lopez, O. C. Lima, B. Andersson, and A. T. de Vasconcelos (2014). “*Trypanosoma cruzi* Clone Dm28c Draft Genome Sequence”. In: *Genome Announc* 2.1. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24482508>.
- Haenni, S., C. K. Renggli, C. M. Fragoso, M. Oberle, and I. Roditi (2006). “The procyclin-associated genes of *Trypanosoma brucei* are not essential for cyclical transmission by tsetse”. In: *Mol Biochem Parasitol* 150.2, pp. 144–56. URL: <https://www.ncbi.nlm.nih.gov/pubmed/16930740>.
- Hafemeister, C. and R. Satija (2019). “Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression”. In: *Genome Biol* 20.1, p. 296. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31870423>.
- Hagemann-Jensen, M., C. Ziegenhain, P. Chen, D. Ramskold, G. J. Hendriks, A. J. M. Larsson, O. R. Faridani, and R. Sandberg (2020). “Single-cell RNA counting at allele and isoform resolution using Smart-seq3”. In: *Nat Biotechnol* 38.6, pp. 708–714. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32518404>.
- Hagemann-Jensen, M., C. Ziegenhain, and R. Sandberg (2022). “Scalable single-cell RNA sequencing from full transcripts with Smart-seq3xpress”. In: *Nat Biotechnol* 40.10, pp. 1452–1457. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35637418>.
- Haghverdi, Laleh, Aaron T. L. Lun, Michael D. Morgan, and John C. Marioni (2018). “Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors”. In: *Nature Biotechnology* 36.5, pp. 421–427.
- Hao, Y., S. Hao, et al. (2021). “Integrated analysis of multimodal single-cell data”. In: *Cell* 184.13, 3573–3587 e29. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34062119>.
- Hao, Y., T. Stuart, M. H. Kowalski, S. Choudhary, P. Hoffman, A. Hartman, A. Srivastava, G. Molla, S. Madad, C. Fernandez-Granda, and R. Satija (2024). “Dictionary learning for integrative, multimodal and scalable single-cell analysis”. In: *Nat Biotechnol* 42.2, pp. 293–304. URL: <https://www.ncbi.nlm.nih.gov/pubmed/37231261>.
- Haque, A., J. Engel, S. A. Teichmann, and T. Lonnberg (2017). “A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications”. In: *Genome Med* 9.1, p. 75. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28821273>.
- Harms, A., E. Maisonneuve, and K. Gerdes (2016). “Mechanisms of bacterial persistence during stress and antibiotic exposure”. In: *Science* 354.6318. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27980159>.
- Hartmann, C., C. Benz, S. Brems, L. Ellis, V. D. Luu, M. Stewart, I. D’Orso, C. Busold, K. Fellenberg, A. C. Frasch, M. Carrington, J. Hoheisel, and C. E. Clayton (2007). “Small trypanosome RNA-binding proteins TbUBP1 and TbUBP2 influence expression of F-box protein mRNAs in bloodstream trypanosomes”. In: *Eukaryot Cell* 6.11, pp. 1964–78. URL: <https://www.ncbi.nlm.nih.gov/pubmed/17873084>.
- Hashimi, H., Z. Cicova, L. Novotna, Y. Z. Wen, and J. Lukes (2009). “Kinetoplastid guide RNA biogenesis is dependent on subunits of the mitochondrial RNA binding complex 1 and mitochondrial RNA polymerase”. In: *RNA* 15.4, pp. 588–99. URL: <https://www.ncbi.nlm.nih.gov/pubmed/19228586>.
- Hashimoto, M., J. Morales, H. Uemura, K. Mikoshiba, and T. Nara (2015). “A Novel Method for Inducing Amastigote-To-Trypomastigote Transformation In Vitro in *Trypanosoma cruzi* Reveals the

- Importance of Inositol 1,4,5-Trisphosphate Receptor”. In: *PLoS One* 10.8, e0135726. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26267656>.
- Hendriks, E. F. and K. R. Matthews (2005). “Disruption of the developmental programme of *Trypanosoma brucei* by genetic ablation of TbZFP1, a differentiation-enriched CCCH protein”. In: *Mol Microbiol* 57.3, pp. 706–16. URL: <https://www.ncbi.nlm.nih.gov/pubmed/16045615>.
- Herreros-Cabello, A., F. Callejas-Hernandez, N. Girones, and M. Fresno (2020). “*Trypanosoma Cruzi* Genome: Organization, Multi-Gene Families, Transcription, and Biological Implications”. In: *Genes (Basel)* 11.10. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33066599>.
- Hirumi, H., J.J. Doyle, and K. Hirumi (1977). “Cultivation of bloodstream *Trypanosoma brucei*”. In: *Bulletin of the World Health Organization* 55.2-3, pp. 405–409.
- Hirumi, H. and K. Hirumi (1989). “Continuous Cultivation of *Trypanosoma brucei* Blood Stream Forms in a Medium Containing a Low Concentration of Serum Protein without Feeder Cell Layers”. In: *The Journal of Parasitology* 75.6, pp. 985–989.
- Hofer-Warbinek, R., J. A. Schmid, H. Mayer, G. Winsauer, L. Orel, B. Mueller, Ch Wiesner, B. R. Binder, and R. de Martin (2004). “A highly conserved proapoptotic gene, IKIP, located next to the APAF1 gene locus, is regulated by p53”. In: *Cell Death Differ* 11.12, pp. 1317–25. URL: <https://www.ncbi.nlm.nih.gov/pubmed/15389287>.
- Howick, V. M., L. Peacock, C. Kay, C. Collett, W. Gibson, and M. K. N. Lawniczak (2022). “Single-cell transcriptomics reveals expression profiles of *Trypanosoma brucei* sexual stages”. In: *PLoS Pathog* 18.3, e1010346. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35255094>.
- Howick, V. M., A. J. C. Russell, T. Andrews, H. Heaton, A. J. Reid, K. Natarajan, H. Butungi, T. Metcalf, L. H. Verzier, J. C. Rayner, M. Berriman, J. K. Herren, O. Billker, M. Hemberg, A. M. Talman, and M. K. N. Lawniczak (2019). “The Malaria Cell Atlas: Single parasite transcriptomes across the complete *Plasmodium* life cycle”. In: *Science* 365.6455. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31439762>.
- Hutchinson, S., S. Foulon, A. Crouzols, R. Menafra, B. Rotureau, A. D. Griffiths, and P. Bastin (2021). “The establishment of variant surface glycoprotein monoallelic expression revealed by single-cell RNA-seq of *Trypanosoma brucei* in the tsetse fly salivary glands”. In: *PLoS Pathog* 17.9, e1009904. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34543350>.
- Hwang, J. R., Y. Byeon, D. Kim, and S. G. Park (2020). “Recent insights of T cell receptor-mediated signaling pathways for T cell activation and development”. In: *Exp Mol Med* 52.5, pp. 750–761. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32439954>.
- Ianevski, A., A. K. Giri, and T. Aittokallio (2022). “Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data”. In: *Nat Commun* 13.1, p. 1246. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35273156>.
- Ilicic, T., J. K. Kim, A. A. Kolodziejczyk, F. O. Bagger, D. J. McCarthy, J. C. Marioni, and S. A. Teichmann (2016). “Classification of low quality cells from single-cell RNA-seq data”. In: *Genome Biol* 17, p. 29. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26887813>.
- Jaccard, P. (1901). “Étude comparative de la distribution florale dans une portion des Alpes et du Jura”. In: *Bulletin de la Societe Vaudoise des Sciences Naturelles* 37, pp. 547–579.
- Jacomy, M., T. Venturini, S. Heymann, and M. Bastian (2014). “ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software”. In: *PLoS One* 9.6, e98679. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24914678>.
- Jiang, R., T. Sun, D. Song, and J. J. Li (2022). “Statistics or biology: the zero-inflation controversy about scRNA-seq data”. In: *Genome Biol* 23.1, p. 31. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35063006>.
- Kafkova, L., C. Tu, K. L. Pazzo, K. P. Smith, E. W. Debler, K. S. Paul, J. Qu, and L. K. Read (2018). “*Trypanosoma brucei* PRMT1 Is a Nucleic Acid Binding Protein with a Role in Energy Metabolism

- and the Starvation Stress Response”. In: *mBio* 9.6. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30563898>.
- Kanehisa, M., M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima (2017). “KEGG: new perspectives on genomes, pathways, diseases and drugs”. In: *Nucleic Acids Res* 45.D1, pp. D353–D361. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27899662>.
- Keren-Shaul, H., E. Kenigsberg, D. A. Jaitin, E. David, F. Paul, A. Tanay, and I. Amit (2019). “MARS-seq2.0: an experimental and analytical pipeline for indexed sorting combined with single-cell RNA sequencing”. In: *Nat Protoc* 14.6, pp. 1841–1862. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31101904>.
- Kessler, R. L., V. T. Contreras, N. P. Marliere, A. Aparecida Guarneri, L. H. Villamizar Silva, Gaca Mazarotto, M. Batista, V. T. Soccol, M. A. Krieger, and C. M. Probst (2017). “Recently differentiated epimastigotes from *Trypanosoma cruzi* are infective to the mammalian host”. In: *Mol Microbiol* 104.5, pp. 712–736. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28240790>.
- Kimuda, M. P., D. Laming, H. C. Hoppe, and O. Tastan Bishop (2019). “Identification of Novel Potential Inhibitors of Pteridine Reductase 1 in *Trypanosoma brucei* via Computational Structure-Based Approaches and in Vitro Inhibition Assays”. In: *Molecules* 24.1. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30609681>.
- Kivioja, T., A. Vaharautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, and J. Taipale (2011). “Counting absolute numbers of molecules using unique molecular identifiers”. In: *Nat Methods* 9.1, pp. 72–4. URL: <https://www.ncbi.nlm.nih.gov/pubmed/22101854>.
- Klein, C., M. Terrao, and C. Clayton (2017). “The role of the zinc finger protein ZC3H32 in bloodstream-form *Trypanosoma brucei*”. In: *PLoS One* 12.5, e0177901. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28545140>.
- Knüsel, S. and I. Roditi (2013). “Insights into the regulation of GPEET procyclin during differentiation from early to late procyclic forms of *Trypanosoma brucei*”. In: *Mol Biochem Parasitol* 191.2, pp. 66–74. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24076427>.
- Koenig-Martin, E., M. Yamage, and I. Roditi (1992). “A procyclin-associated gene in *Trypanosoma brucei* encodes a polypeptide related to ESAG 6 and 7 proteins”. In: *Molecular and Biochemical Parasitology* 55, pp. 135–146.
- Kolev, N. G., K. Ramey-Butler, G. A. M. Cross, E. Ullu, and C. Tschudi (2012). “Developmental progression to infectivity in *Trypanosoma brucei* triggered by an RNA-binding protein”. In: *Science* 338.6112, pp. 1352–3. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23224556>.
- Kolev, N. G., C. Tschudi, and E. Ullu (2011). “RNA interference in protozoan parasites: achievements and challenges”. In: *Eukaryot Cell* 10.9, pp. 1156–63. URL: <https://www.ncbi.nlm.nih.gov/pubmed/21764910>.
- Korsunsky, Ilya (2020). *About the convergence plot*. Web Page. URL: <https://github.com/immunogenomics/harmony/issues/67>.
- Korsunsky, Ilya, Aparna Nathan, Nghia Millard, and Soumya Raychaudhuri (2019). “Presto scales Wilcoxon and auROC analyses to millions of observations”. In: *BioRxiv*.
- Kreslavsky, T., M. Gleimer, M. Miyazaki, Y. Choi, E. Gagnon, C. Murre, P. Sicinski, and H. von Boehmer (2012). “beta-Selection-induced proliferation is required for alphabeta T cell differentiation”. In: *Immunity* 37.5, pp. 840–53. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23159226>.
- Kulkarni, M. M., C. L. Olson, D. M. Engman, and B. S. McGwire (2009). “*Trypanosoma cruzi* GP63 proteins undergo stage-specific differential posttranslational modification and are important for host cell infection”. In: *Infect Immun* 77.5, pp. 2193–200. URL: <https://www.ncbi.nlm.nih.gov/pubmed/19273559>.
- La Manno, G. et al. (2018). “RNA velocity of single cells”. In: *Nature* 560.7719, pp. 494–498. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30089906>.

- Laidlaw, Ross F., Emma M. Briggs, Keith R. Matthews, Richard McCulloch, and Thomas D. Otto (2024). “TrAGEDy: Trajectory Alignment of Gene Expression Dynamics”. In: *BioRxiv*.
- Lamour, N., L. Riviere, V. Coustou, G. H. Coombs, M. P. Barrett, and F. Bringaud (2005). “Proline metabolism in procyclic *Trypanosoma brucei* is down-regulated in the presence of glucose”. In: *J Biol Chem* 280.12, pp. 11902–10. URL: <https://www.ncbi.nlm.nih.gov/pubmed/15665328>.
- Lange, M., V. Bergen, M. Klein, M. Setty, B. Reuter, M. Bakhti, H. Lickert, M. Ansari, J. Schniering, H. B. Schiller, D. Pe’er, and F. J. Theis (2022). “CellRank for directed single-cell fate mapping”. In: *Nat Methods* 19.2, pp. 159–170. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35027767>.
- Levsky, J.M., S.M. Shenoy, R.C. Pezo, and R.H. Singer (2002). “Single-Cell Gene Expression Profiling”. In: *Science* 297.
- Ley, A., E.S. Robbins, V. Nussenzweig, and N.W. Andrews (1990). “The exit of trypanosoma cruzi from the phagosome is inhibited by raising the pH of acidic compartments”. In: *Journal of Experimental Medicine* 171, pp. 401–413.
- Li, G. Q., J. Xia, W. Zeng, W. Luo, L. Liu, X. Zeng, and D. Cao (2023). “The intestinal gammadelta T cells: functions in the gut and in the distant organs”. In: *Front Immunol* 14, p. 1206299. URL: <https://www.ncbi.nlm.nih.gov/pubmed/37398661>.
- Li, S. et al. (2014). “Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study”. In: *Nat Biotechnol* 32.9, pp. 915–925. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25150835>.
- Li, Y., S. Shah-Simpson, K. Okrah, A. T. Belew, J. Choi, K. L. Caradonna, P. Padmanabhan, D. M. Ndegwa, M. R. Temanni, H. Corrada Bravo, N. M. El-Sayed, and B. A. Burleigh (2016). “Transcriptome Remodeling in *Trypanosoma cruzi* and Human Cells during Intracellular Infection”. In: *PLoS Pathog* 12.4, e1005511. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27046031>.
- Liao, Y., G. K. Smyth, and W. Shi (2014). “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features”. In: *Bioinformatics* 30.7, pp. 923–30. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24227677>.
- Ling, A. S., J. R. Trotter, and E. F. Hendriks (2011). “A zinc finger protein, TbZC3H20, stabilizes two developmentally regulated mRNAs in trypanosomes”. In: *J Biol Chem* 286.23, pp. 20152–62. URL: <https://www.ncbi.nlm.nih.gov/pubmed/21467035>.
- Linsley, P.S., J.L. Greene, P. Tan, J. Bradshaw, J.A. Ledbetter, C. Anasetti, and N.K. Damle (1992). “Coexpression and Functional Cooperation of CTLA-4 and CD28 on Activated T Lymphocytes”. In: *Journal of Experimental Medicine* 176, pp. 1595–1604.
- Liu, B. and C. Clayton (2022). “Gel shift experiments with fragments of the *Trypanosoma brucei* RNA-binding protein RBP10”. In: *BMC Res Notes* 15.1, p. 253. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35841065>.
- Liu, B., K. Kamanyi Marucha, and C. Clayton (2020). “The zinc finger proteins ZC3H20 and ZC3H21 stabilise mRNAs encoding membrane proteins and mitochondrial proteins in insect-form *Trypanosoma brucei*”. In: *Mol Microbiol* 113.2, pp. 430–451. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31743541>.
- Liu, L., Y. X. Xu, K. L. Caradonna, E. K. Kruzel, B. A. Burleigh, J. D. Bangs, and C. B. Hirschberg (2013). “Inhibition of nucleotide sugar transport in *Trypanosoma brucei* alters surface glycosylation”. In: *J Biol Chem* 288.15, pp. 10599–615. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23443657>.
- Louradour, I., T. R. Ferreira, E. Duge, N. Karunaweera, A. Paun, and D. Sacks (2022). “Stress conditions promote *Leishmania* hybridization in vitro marked by expression of the ancestral gamete fusogen HAP2 as revealed by single-cell RNA-seq”. In: *Elife* 11. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34994687>.
- Love, M. I., W. Huber, and S. Anders (2014). “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome Biol* 15.12, p. 550. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25516281>.

- Low, H. P., J. J. Paulin, and C. H. Keith (1992). “Trypanosoma cruzi infection of BSC-1 fibroblast cells causes cytoskeletal disruption and changes in intracellular calcium levels”. In: *J Protozool* 39.4, pp. 463–70. URL: <https://www.ncbi.nlm.nih.gov/pubmed/1403981>.
- Luecken, M. D., M. Buttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colome-Tatche, and F. J. Theis (2022). “Benchmarking atlas-level data integration in single-cell genomics”. In: *Nat Methods* 19.1, pp. 41–50. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34949812>.
- Luecken, M. D. and F. J. Theis (2019). “Current best practices in single-cell RNA-seq analysis: a tutorial”. In: *Mol Syst Biol* 15.6, e8746. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31217225>.
- Lueong, S., C. Merce, B. Fischer, J. D. Hoheisel, and E. D. Erben (2016). “Gene expression regulatory networks in Trypanosoma brucei: insights into the role of the mRNA-binding proteome”. In: *Mol Microbiol* 100.3, pp. 457–71. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26784394>.
- Lye, L. F., K. L. Owens, S. Jang, J. E. Marcus, E. A. Brettmann, and S. M. Beverley (2022). “An RNA Interference (RNAi) Toolkit and Its Utility for Functional Genetic Analysis of Leishmania (Viannia)”. In: *Genes (Basel)* 14.1. URL: <https://www.ncbi.nlm.nih.gov/pubmed/36672832>.
- Maas-Bauer, K., N. Kohler, A. V. Stell, M. Zwick, S. Acharya, A. Rensing-Ehl, C. Konig, J. Kroll, J. Baker, S. Kossmann, A. Pradier, S. Wang, M. Docquier, D. B. Lewis, R. S. Negrin, and F. Simonetta (2024). “Single-cell transcriptomics reveal different maturation stages and sublineage commitment of human thymic invariant natural killer T cells”. In: *J Leukoc Biol* 115.2, pp. 401–409. URL: <https://www.ncbi.nlm.nih.gov/pubmed/37742056>.
- Maaten, L. van der and G. Hinton (2008). “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 1, pp. 1–48.
- MacLean, L. M., J. Thomas, M. D. Lewis, I. Cotillo, D. W. Gray, and M. De Rycker (2018). “Development of Trypanosoma cruzi in vitro assays to identify compounds suitable for progression in Chagas’ disease drug discovery”. In: *PLoS Negl Trop Dis* 12.7, e0006612. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30001347>.
- Macosko, E. Z., A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll (2015). “Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets”. In: *Cell* 161.5, pp. 1202–1214. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26000488>.
- Madeira, F., N. Madhusoodanan, J. Lee, A. Eusebi, A. Niewielska, A. R. N. Tivey, R. Lopez, and S. Butcher (2024). “The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024”. In: *Nucleic Acids Res* 52.W1, W521–W525. URL: <https://www.ncbi.nlm.nih.gov/pubmed/38597606>.
- Maizels, R. J., D. M. Snell, and J. Briscoe (2024). “Reconstructing developmental trajectories using latent dynamical systems and time-resolved transcriptomics”. In: *Cell Syst* 15.5, 411–424 e9. URL: <https://www.ncbi.nlm.nih.gov/pubmed/38754365>.
- Manful, T., A. Fadda, and C. Clayton (2011). “The role of the 5’-3’ exoribonuclease XRNA in transcriptome-wide mRNA degradation”. In: *RNA* 17.11, pp. 2039–47. URL: <https://www.ncbi.nlm.nih.gov/pubmed/21947264>.
- Mantilla, B. S., L. Marchese, A. Casas-Sanchez, N. A. Dyer, N. Ejeh, M. Biran, F. Bringaud, M. J. Lehane, A. Acosta-Serrano, and A. M. Silber (2017). “Proline Metabolism is Essential for Trypanosoma brucei Survival in the Tsetse Vector”. In: *PLoS Pathog* 13.1, e1006158. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28114403>.
- Marinov, G. K., B. A. Williams, K. McCue, G. P. Schroth, J. Gertz, R. M. Myers, and B. J. Wold (2014). “From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing”. In: *Genome Res* 24.3, pp. 496–510. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24299736>.
- Martin-Escolano, J., C. Marin, M. J. Rosales, A. D. Tsaousis, E. Medina-Carmona, and R. Martin-Escolano (2022). “An Updated View of the Trypanosoma cruzi Life Cycle: Intervention Points for an

- Effective Treatment”. In: *ACS Infect Dis* 8.6, pp. 1107–1115. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35652513>.
- Martins, R. M., R. M. Alves, S. Macedo, and N. Yoshida (2011). “Starvation and rapamycin differentially regulate host cell lysosome exocytosis and invasion by *Trypanosoma cruzi* metacyclic forms”. In: *Cell Microbiol* 13.7, pp. 943–54. URL: <https://www.ncbi.nlm.nih.gov/pubmed/21501360>.
- Matthews, K. R. (2005). “The developmental cell biology of *Trypanosoma brucei*”. In: *J Cell Sci* 118.Pt 2, pp. 283–90. URL: <https://www.ncbi.nlm.nih.gov/pubmed/15654017>.
- (2021). “Trypanosome Signaling-Quorum Sensing”. In: *Annu Rev Microbiol* 75, pp. 495–514. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34348028>.
- Matthews, K. R., J. R. Ellis, and A. Paterou (2004). “Molecular regulation of the life cycle of African trypanosomes”. In: *Trends Parasitol* 20.1, pp. 40–7. URL: <https://www.ncbi.nlm.nih.gov/pubmed/14700589>.
- McInnes, Leland, John Healy, Nathaniel Saul, and Lukas Großberger (2018). “UMAP: Uniform Manifold Approximation and Projection”. In: *Journal of Open Source Software* 3.29.
- Minning, T. A., D. B. Weatherly, 3rd Atwood J., R. Orlando, and R. L. Tarleton (2009). “The steady-state transcriptome of the four major life-cycle stages of *Trypanosoma cruzi*”. In: *BMC Genomics* 10, p. 370. URL: <https://www.ncbi.nlm.nih.gov/pubmed/19664227>.
- Miranda, M. R., G. E. Canepa, L. A. Bouvier, and C. A. Pereira (2006). “*Trypanosoma cruzi*: Oxidative stress induces arginine kinase expression”. In: *Exp Parasitol* 114.4, pp. 341–4. URL: <https://www.ncbi.nlm.nih.gov/pubmed/16725140>.
- Mombaerts, P., J. Iacomini, R.S. Johnson, K. Herrup, S. Tonegawa, and V.E. Papaioannou (1992). “RAG-1-Deficient Mice Have No Mature 8 and T Lymphocytes”. In: *Cell* 68, pp. 869–877.
- Moon, K. R., D. van Dijk, Z. Wang, S. Gigante, D. B. Burkhardt, W. S. Chen, K. Yim, A. V. D. Elzen, M. J. Hirn, R. R. Coifman, N. B. Ivanova, G. Wolf, and S. Krishnaswamy (2019). “Visualizing structure and transitions in high-dimensional biological data”. In: *Nat Biotechnol* 37.12, pp. 1482–1492. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31796933>.
- Mugo, E. and C. Clayton (2017). “Expression of the RNA-binding protein RBP10 promotes the bloodstream-form differentiation state in *Trypanosoma brucei*”. In: *PLoS Pathog* 13.8, e1006560. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28800584>.
- Müller, L. S. M., R. O. Cosentino, K. U. Forstner, J. Guizetti, C. Wedel, N. Kaplan, C. J. Janzen, P. Arampatzi, J. Vogel, S. Steinbiss, T. D. Otto, A. E. Saliba, R. P. Sebra, and T. N. Siegel (2018). “Genome organization and DNA accessibility control antigenic variation in trypanosomes”. In: *Nature* 563.7729, pp. 121–125. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30333624>.
- Naguleswaran, A., P. Fernandes, S. Bevkal, R. Rehmann, P. Nicholson, and I. Roditi (2021). “Developmental changes and metabolic reprogramming during establishment of infection and progression of *Trypanosoma brucei brucei* through its insect host”. In: *PLoS Negl Trop Dis* 15.9, e0009504. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34543277>.
- Nardy, A.F., C. G. Freire-de-Lima, and A. Morrot (2015). “Immune Evasion Strategies of *Trypanosoma cruzi*”. In: *J Immunol Res* 2015, p. 178947. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26240832>.
- Ng, S. S. et al. (2020). “The NK cell granule protein NKG7 regulates cytotoxic granule exocytosis and inflammation”. In: *Nat Immunol* 21.10, pp. 1205–1218. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32839608>.
- Ngoune, J.M.T., P. Sharma, A. Crouzols, and B. Rotureau (2024). “Stumpy forms are the predominant transmissible forms of *Trypanosoma brucei*”. In: *eLife*.
- Oberle, M., O. Balmer, R. Brun, and I. Roditi (2010). “Bottlenecks and the maintenance of minor genotypes during the life cycle of *Trypanosoma brucei*”. In: *PLoS Pathog* 6.7, e1001023. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20686656>.

- Oetjen, K. A., K. E. Lindblad, M. Goswami, G. Gui, P. K. Dagur, C. Lai, L. W. Dillon, J. P. McCoy, and C. S. Hourigan (2018). “Human bone marrow assessment by single-cell RNA sequencing, mass cytometry, and flow cytometry”. In: *JCI Insight* 3.23. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30518681>.
- Ooi, C. P., B. Rotureau, S. Gribaldo, C. Georgikou, D. Julkowska, T. Blisnick, S. Perrot, I. Subota, and P. Bastin (2015). “The Flagellar Arginine Kinase in *Trypanosoma brucei* Is Important for Infection in Tsetse Flies”. In: *PLoS One* 10.7, e0133676. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26218532>.
- Parekh, S., C. Ziegenhain, B. Vieth, W. Enard, and I. Hellmann (2016). “The impact of amplification on differential expression analyses by RNA-seq”. In: *Sci Rep* 6, p. 25533. URL: <https://www.ncbi.nlm.nih.gov/pubmed/27156886>.
- Pearson, Karl (1901). “On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11, pp. 559–572.
- Pereira, M. E. A., K. Zhang, Y. Gong, E.M. Herrera, and M. Ming (1996). “Invasive Phenotype of *Trypanosoma cruzi* Restricted to a Population Expressing trans-Sialidase”. In: *Infection and Immunity* 64.9, pp. 3884–3892.
- Picelli, S., O. R. Faridani, A. K. Bjorklund, G. Winberg, S. Sagasser, and R. Sandberg (2014). “Full-length RNA-seq from single cells using Smart-seq2”. In: *Nat Protoc* 9.1, pp. 171–81. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24385147>.
- Piras, M.M (1982). “Changes in morphology and infectivity of cell culture-derived trypomastigotes of *trypanosoma cruzi*”. In: *Molecular and Biochemical Parasitology* 6, pp. 67–81.
- Pita, S., F. Diaz-Viraque, G. Iraola, and C. Robello (2019). “The Tritryps Comparative Repeatome: Insights on Repetitive Element Evolution in Trypanosomatid Pathogens”. In: *Genome Biol Evol* 11.2, pp. 546–551. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30715360>.
- Polanski, K., M. D. Young, Z. Miao, K. B. Meyer, S. A. Teichmann, and J. E. Park (2020). “BBKNN: fast batch alignment of single cell transcriptomes”. In: *Bioinformatics* 36.3, pp. 964–965. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31400197>.
- Poran, A., C. Notzel, O. Aly, N. Mencia-Trinchant, C. T. Harris, M. L. Guzman, D. C. Hassane, O. Elemento, and B. F. C. Kafack (2017). “Single-cell RNA sequencing reveals a signature of sexual commitment in malaria parasites”. In: *Nature* 551.7678, pp. 95–99. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29094698>.
- Probst, V., A. Simonyan, F. Pacheco, Y. Guo, F. C. Nielsen, and F. O. Bagger (2022). “Benchmarking full-length transcript single cell mRNA sequencing protocols”. In: *BMC Genomics* 23.1, p. 860. URL: <https://www.ncbi.nlm.nih.gov/pubmed/36581800>.
- Qiu, P. (2020). “Embracing the dropouts in single-cell RNA-seq analysis”. In: *Nat Commun* 11.1, p. 1169. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32127540>.
- Qiu, X., A. Hill, J. Packer, D. Lin, Y. A. Ma, and C. Trapnell (2017). “Single-cell mRNA quantification and differential analysis with Census”. In: *Nat Methods* 14.3, pp. 309–315. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28114287>.
- Qiu, Y., J. E. Milanes, J. A. Jones, R. E. Noorai, V. Shankar, and J. C. Morris (2018). “Glucose Signaling Is Important for Nutrient Adaptation during Differentiation of Pleomorphic African Trypanosomes”. In: *mSphere* 3.5. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30381351>.
- Ralph, Stuart Alexander, Guillermo Bernabó, Gabriela Levy, María Ziliani, Lucas D. Caeiro, Daniel O. Sánchez, and Valeria Tekiel (2013). “TcTASV-C, a Protein Family in *Trypanosoma cruzi* that Is Predominantly Trypomastigote-Stage Specific and Secreted to the Medium”. In: *PLoS ONE* 8.7.
- Ramirez, M.I., R. de Cassia Ruiz, J.E. Araya, J.F. da Silveira, and N. Yoshida (1993). “Involvement of the Stage-Specific 82-Kilodalton Adhesion Molecule of *Trypanosoma cruzi* Metacyclic Trypomastigotes in Host Cell Invasion”. In: *Infection and Immunity* 61, pp. 3636–3641.

- Rangel-Aldao, R., O. Allende, F. Triana, R. Piras, D. Henriquez, and M. Piras (1987). “Possible role of cAMP in the differentiation of *Trypanosoma cruzi*”. In: *Molecular and Biochemical Parasitology* 22, pp. 39–43.
- Real, E., V. M. Howick, F. A. Dahalan, K. Witmer, J. Cudini, C. Andradi-Brown, J. Blight, M. S. Davidson, S. K. Dogga, A. J. Reid, J. Baum, and M. K. N. Lawniczak (2021). “A single-cell atlas of *Plasmodium falciparum* transmission through the mosquito”. In: *Nat Commun* 12.1, p. 3196. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34045457>.
- Reid, A. J., A. M. Talman, H. M. Bennett, A. R. Gomes, M. J. Sanders, C. J. R. Illingworth, O. Billker, M. Berriman, and M. K. Lawniczak (2018). “Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites”. In: *Elife* 7. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29580379>.
- Reina-San-Martín, B., W. Degraeve, C. Rougeot, A. Cosson, N. Chamond, A. Cordeiro-Da-Silva, M. Arala-Chaves, A. Coutinho, and P. Minoprio (2000). “A B-cell mitogen from a pathogenic trypanosome is a eukaryotic proline racemase”. In: *Nature Medicine* 6.8, pp. 890–897.
- Resende, B. C., A. C. S. Oliveira, A. C. P. Guanabens, B. M. Repoles, V. Santana, P. M. Hiraiwa, S. D. J. Pena, G. R. Franco, A. M. Macedo, E. B. Tahara, S. P. Fragoso, L. O. Andrade, and C. R. Machado (2020). “The Influence of Recombinational Processes to Induce Dormancy in *Trypanosoma cruzi*”. In: *Front Cell Infect Microbiol* 10, p. 5. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32117793>.
- Reuner, B., E. Vassella, B. Yutzy, and M. Boshart (1997). “Cell density triggers slender to stumpy differentiation of *Trypanosoma brucei* bloodstream forms in culture”. In: *Molecular and Biochemical Parasitology* 90.1, pp. 269–80.
- Rhee, S. G., H. A. Woo, I. S. Kil, and S. H. Bae (2012). “Peroxiredoxin functions as a peroxidase and a regulator and sensor of local peroxides”. In: *J Biol Chem* 287.7, pp. 4403–10. URL: <https://www.ncbi.nlm.nih.gov/pubmed/22147704>.
- Rico, E., L. Jeacock, J. Kovarova, and D. Horn (2018). “Inducible high-efficiency CRISPR-Cas9-targeted gene editing and precision base editing in African trypanosomes”. In: *Sci Rep* 8.1, p. 7960. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29785042>.
- Rico, E., F. Rojas, B. M. Mony, B. Szoor, P. Macgregor, and K. R. Matthews (2013). “Bloodstream form pre-adaptation to the tsetse fly in *Trypanosoma brucei*”. In: *Front Cell Infect Microbiol* 3, p. 78. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24294594>.
- Roberts, A. J., L. S. Torrie, S. Wyllie, and A. H. Fairlamb (2014). “Biochemical and genetic characterization of *Trypanosoma cruzi* N-myristoyltransferase”. In: *Biochem J* 459.2, pp. 323–32. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24444291>.
- Rodriguez-Bejarano, O. H., C. Avendano, and M. A. Patarroyo (2021). “Mechanisms Associated with *Trypanosoma cruzi* Host Target Cell Adhesion, Recognition and Internalization”. In: *Life (Basel)* 11.6. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34207491>.
- Rojas, F., E. Silvester, J. Young, R. Milne, M. Tettey, D. R. Houston, M. D. Walkinshaw, I. Perez-Pi, M. Auer, H. Denton, T. K. Smith, J. Thompson, and K. R. Matthews (2019). “Oligopeptide Signaling through TbGPR89 Drives Trypanosome Quorum Sensing”. In: *Cell* 176.1-2, 306–317 e16. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30503212>.
- Romagnoli, B. A. A., F. B. Holetz, L. R. Alves, and S. Goldenberg (2020). “RNA Binding Proteins and Gene Expression Regulation in *Trypanosoma cruzi*”. In: *Front Cell Infect Microbiol* 10, p. 56. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32154189>.
- Romaniuk, M. A., G. Cervini, and A. Cassola (2016). “Regulation of RNA binding proteins in trypanosomatid protozoan parasites”. In: *World J Biol Chem* 7.1, pp. 146–57. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26981203>.
- Rotureau, B., I. Subota, J. Buisson, and P. Bastin (2012). “A new asymmetric division contributes to the continuous production of infective trypanosomes in the tsetse fly”. In: *Development* 139.10, pp. 1842–50. URL: <https://www.ncbi.nlm.nih.gov/pubmed/22491946>.

- Rotureau, B. and J. Van Den Abbeele (2013). “Through the dark continent: African trypanosome development in the tsetse fly”. In: *Front Cell Infect Microbiol* 3, p. 53. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24066283>.
- Ruiz, R.C., S. Favoreto Jr, M.L. Dorta, M.E.M. Oshiro, A.T. Ferreira, P.M. Manque, and N. Yoshida (1998). “Infectivity of *Trypanosoma cruzi* strains is associated with differential expression of surface glycoproteins with differential Ca²⁺ signalling activity”. In: *330*, pp. 505–511.
- Sabalette, K. B., J. R. Sotelo-Silveira, P. Smircich, and J. G. De Gaudenzi (2023). “RNA-Seq reveals that overexpression of TcUBP1 switches the gene expression pattern toward that of the infective form of *Trypanosoma cruzi*”. In: *J Biol Chem* 299.5, p. 104623. URL: <https://www.ncbi.nlm.nih.gov/pubmed/36935010>.
- Saelens, W., R. Cannoodt, H. Todorov, and Y. Saeys (2019). “A comparison of single-cell trajectory inference methods”. In: *Nat Biotechnol* 37.5, pp. 547–554. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30936559>.
- Sánchez-Valdéz, F. J., A. Padilla, W. Wang, D. Orr, and R. L. Tarleton (2018). “Spontaneous dormancy protects *Trypanosoma cruzi* during extended drug exposure”. In: *Elife* 7. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29578409>.
- Sansom, D.M. (2000). “CD28, CTLA-4 and their ligands: who does what and to whom?” In: *Immunology* 101, pp. 169–177.
- Santos, C. C., C. Sant’anna, A. Terres, N. L. Cunha-e-Silva, J. Scharfstein, and A. Lima A. P. de (2005). “Chagasin, the endogenous cysteine-protease inhibitor of *Trypanosoma cruzi*, modulates parasite differentiation and invasion of mammalian cells”. In: *J Cell Sci* 118.Pt 5, pp. 901–15. URL: <https://www.ncbi.nlm.nih.gov/pubmed/15713748>.
- Santos, Cmbd, A. Ludwig, R. L. Kessler, R. C. P. Rampazzo, A. H. Inoue, M. A. Krieger, D. P. Pavoni, and C. M. Probst (2018). “*Trypanosoma cruzi* transcriptome during axenic epimastigote growth curve”. In: *Mem Inst Oswaldo Cruz* 113.5, e170404. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29668769>.
- Santos, V. C., A. E. R. Oliveira, A. C. B. Campos, J. L. Reis-Cunha, D. C. Bartholomeu, S. M. R. Teixeira, A. Lima, and R. S. Ferreira (2021). “The gene repertoire of the main cysteine protease of *Trypanosoma cruzi*, cruzipain, reveals four sub-types with distinct active sites”. In: *Sci Rep* 11.1, p. 18231. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34521898>.
- Satija, R., J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev (2015). “Spatial reconstruction of single-cell gene expression data”. In: *Nat Biotechnol* 33.5, pp. 495–502. URL: <https://www.ncbi.nlm.nih.gov/pubmed/25867923>.
- Savage, A. F., N. G. Kolev, J. B. Franklin, A. Vigneron, S. Aksoy, and C. Tschudi (2016). “Transcriptome Profiling of *Trypanosoma brucei* Development in the Tsetse Fly Vector *Glossina morsitans*”. In: *PLoS One* 11.12, e0168877. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28002435>.
- Sbicego, S., J. D. Alfonzo, A. M. Estevez, M. A. Rubio, X. Kang, C. W. Turck, M. Peris, and L. Simpson (2003). “RBP38, a novel RNA-binding protein from trypanosomatid mitochondria, modulates RNA stability”. In: *Eukaryot Cell* 2.3, pp. 560–8. URL: <https://www.ncbi.nlm.nih.gov/pubmed/12796301>.
- Scanpy (2024). *Preprocessing and clustering*. Web Page. URL: <https://scanpy.readthedocs.io/en/stable/tutorials/basics/clustering.html>.
- scanpy (n.d.). *Integrating data using ingest and BBKNN*. Web Page.
- Schindelin, J., I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J. Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona (2012). “Fiji: an open-source platform for biological-image analysis”. In: *Nat Methods* 9.7, pp. 676–82. URL: <https://www.ncbi.nlm.nih.gov/pubmed/22743772>.

- Schmatz, D. M., R. C. Boltz, and P. K. Murray (1983). “Trypanosoma cruzi: separation of broad and slender trypomastigotes using a continuous hypaque gradient”. In: *Parasitology* 87 (Pt 2), pp. 219–27. URL: <https://www.ncbi.nlm.nih.gov/pubmed/6359027>.
- Schmidt, A., J. Sivaraman, Y. Li, R. Larocque, J. A. R. G. Barbosa, C. Smith, A. Matte, J.D. Schrag, and M. Cygler (2001). “Three-Dimensional Structure of 2-Amino-3-ketobutyrate CoA Ligase from Escherichia coli Complexed with a PLP-Substrate Intermediate: Inferred Reaction Mechanism”. In: *Biochemistry* 40, pp. 5151–5160.
- Schönenberger, M. and R. Brun (1979). “Cultivation and in vitro cloning of procyclic culture forms of ”Trypanosoma brucei” in a semi-defined medium: short communication”. In: *Acta Tropica* 36.
- Schraivogel, D., A. R. Gschwind, J. H. Milbank, D. R. Leonce, P. Jakob, L. Mathur, J. O. Korbel, C. A. Merten, L. Velten, and L. M. Steinmetz (2020). “Targeted Perturb-seq enables genome-scale genetic screens in single cells”. In: *Nat Methods* 17.6, pp. 629–635. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32483332>.
- Schuster, S., J. Lisack, I. Subota, H. Zimmermann, C. Reuter, T. Mueller, B. Morriswood, and M. Engstler (2021). “Unexpected plasticity in the life cycle of Trypanosoma brucei”. In: *Elife* 10. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34355698>.
- Seurat (2023). “Integrative analysis in Seurat v5”. In: URL: https://satijalab.org/seurat/articles/seurat5_integration.
- Shah-Simpson, S., G. Lentini, P. C. Dumoulin, and B. A. Burleigh (2017). “Modulation of host central carbon metabolism and in situ glucose uptake by intracellular Trypanosoma cruzi amastigotes”. In: *PLoS Pathog* 13.11, e1006747. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29176805>.
- Shanmugasundram, A., D. Starns, U. Bohme, B. Amos, P. A. Wilkinson, O. S. Harb, S. Warrenfeltz, J. C. Kissinger, M. A. McDowell, D. S. Roos, K. Crouch, and A. R. Jones (2023). “TriTrypDB: An integrated functional genomics resource for kinetoplastida”. In: *PLoS Negl Trop Dis* 17.1, e0011058. URL: <https://www.ncbi.nlm.nih.gov/pubmed/36656904>.
- Sharma, R., E. Gluenz, L. Peacock, W. Gibson, K. Gull, and M. Carrington (2009). “The heart of darkness: growth and form of Trypanosoma brucei in the tsetse fly”. In: *Trends Parasitol* 25.11, pp. 517–24. URL: <https://www.ncbi.nlm.nih.gov/pubmed/19747880>.
- Shi, H., K. Butler, and C. Tschudi (2018). “A single-point mutation in the RNA-binding protein 6 generates Trypanosoma brucei metacyclics that are able to progress to bloodstream forms in vitro”. In: *Mol Biochem Parasitol* 224, pp. 50–56. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30055184>.
- Shi, H., Y. Zhou, E. Jia, M. Pan, Y. Bai, and Q. Ge (2021). “Bias in RNA-seq Library Preparation: Current Challenges and Solutions”. In: *Biomed Res Int* 2021, p. 6647597. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33987443>.
- Siegel, T. N., D. R. Hekstra, X. Wang, S. Dewell, and G. A. Cross (2010). “Genome-wide analysis of mRNA abundance in two life-cycle stages of Trypanosoma brucei and identification of splicing and polyadenylation sites”. In: *Nucleic Acids Res* 38.15, pp. 4946–57. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20385579>.
- Silva Dias Vieira, C., R. Pinheiro Aguiar, N. P. de Almeida Nogueira, G. Costa Dos Santos Junior, and M. C. Paes (2023). “Glucose metabolism sustains heme-induced Trypanosoma cruzi epimastigote growth in vitro”. In: *PLoS Negl Trop Dis* 17.11, e0011725. URL: <https://www.ncbi.nlm.nih.gov/pubmed/37948458>.
- Smircich, P., G. Eastman, S. Bispo, M. A. Duhagon, E. P. Guerra-Slombo, B. Garat, S. Goldenberg, D. J. Munroe, B. Dallagiovanna, F. Holetz, and J. R. Sotelo-Silveira (2015). “Ribosome profiling reveals translation control as a key mechanism generating differential gene expression in Trypanosoma cruzi”. In: *BMC Genomics* 16.1, p. 443. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26054634>.
- Smircich, P., L. Perez-Diaz, F. Hernandez, M. A. Duhagon, and B. Garat (2023). “Transcriptomic analysis of the adaptation to prolonged starvation of the insect-dwelling Trypanosoma cruzi epimastigotes”.

- In: *Front Cell Infect Microbiol* 13, p. 1138456. URL: <https://www.ncbi.nlm.nih.gov/pubmed/37091675>.
- Song, D. and J. J. Li (2021). “PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell RNA sequencing data”. In: *Genome Biol* 22.1, p. 124. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33926517>.
- Stecconi-Silva, R.B., W.K. Andreoli, and R.A. Mortara (2003). “Parameters Affecting Cellular Invasion and Escape from the Parasitophorous Vacuole by Different Infective Forms of *Trypanosoma cruzi*”. In: *Memórias do Instituto Oswaldo Cruz* 98.7, pp. 953–958.
- Street, K., D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, and S. Dudoit (2018). “Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics”. In: *BMC Genomics* 19.1, p. 477. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29914354>.
- Stuart, K., R. Brun, S. Croft, A. Fairlamb, R. E. Gurtler, J. McKerrow, S. Reed, and R. Tarleton (2008). “Kinetoplastids: related protozoan pathogens, different diseases”. In: *J Clin Invest* 118.4, pp. 1301–10. URL: <https://www.ncbi.nlm.nih.gov/pubmed/18382742>.
- Stuart, T., A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, 3rd Mauck W. M., Y. Hao, M. Stoeckius, P. Smibert, and R. Satija (2019). “Comprehensive Integration of Single-Cell Data”. In: *Cell* 177.7, 1888–1902 e21. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31178118>.
- Sumanaweera, Dinithi, Chenqu Suo, Ana-Maria Cujba, Daniele Muraro, Emma Dann, Krzysztof Polanski, Alexander S. Steemers, Wochan Lee, Amanda J. Oliver, Jong-Eun Park, Kerstin B. Meyer, Bianca Dumitrascu, and Sarah A. Teichmann (2023). “Gene-level alignment of single cell trajectories”. In: *bioRxiv*.
- Sun, Z., L. Chen, H. Xin, Y. Jiang, Q. Huang, A. R. Cillo, T. Tabib, J. K. Kolls, T. C. Bruno, R. Lafyatis, D. A. A. Vignali, K. Chen, Y. Ding, M. Hu, and W. Chen (2019). “A Bayesian mixture model for clustering droplet-based single-cell transcriptomic data from population studies”. In: *Nat Commun* 10.1, p. 1649. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30967541>.
- Sunter, J. D. and K. Gull (2016). “The Flagellum Attachment Zone: ‘The Cellular Ruler’ of Trypanosome Morphology”. In: *Trends Parasitol* 32.4, pp. 309–324. URL: <https://www.ncbi.nlm.nih.gov/pubmed/26776656>.
- Tabula Muris, Consortium, coordination Overall, coordination Logistical, collection Organ, processing, preparation Library, sequencing, analysis Computational data, annotation Cell type, group Writing, group Supplemental text writing, and investigators Principal (2018). “Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris”. In: *Nature* 562.7727, pp. 367–372. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30283141>.
- Tabula Sapiens, Consortium et al. (2022). “The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans”. In: *Science* 376.6594, eabl4896. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35549404>.
- Taleva, G., M. Husova, B. Panicucci, C. Hierro-Yap, E. Pineda, M. Biran, M. Moos, P. Simek, F. Butter, F. Bringaud, and A. Zikova (2023). “Mitochondrion of the *Trypanosoma brucei* long slender bloodstream form is capable of ATP production by substrate-level phosphorylation”. In: *PLoS Pathog* 19.10, e1011699. URL: <https://www.ncbi.nlm.nih.gov/pubmed/37819951>.
- Tang, F., C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani (2009). “mRNA-Seq whole-transcriptome analysis of a single cell”. In: *Nat Methods* 6.5, pp. 377–82. URL: <https://www.ncbi.nlm.nih.gov/pubmed/19349980>.
- Tavares, T. S., F. L. B. Mugge, V. Grazielle-Silva, B. M. Valente, W. M. Goes, A. E. R. Oliveira, A. T. Belew, A. A. Guarneri, F. S. Pais, N. M. El-Sayed, and S. M. R. Teixeira (2021). “A *Trypanosoma cruzi* zinc finger protein that is implicated in the control of epimastigote-specific gene expression and metacyclogenesis”. In: *Parasitology* 148.10, pp. 1171–1185. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33190649>.

- Taylor, M. C., A. Ward, F. Olmo, S. Jayawardhana, A. F. Francisco, M. D. Lewis, and J. M. Kelly (2020). “Intracellular DNA replication and differentiation of *Trypanosoma cruzi* is asynchronous within individual host cells in vivo at all stages of infection”. In: *PLoS Negl Trop Dis* 14.3, e0008007. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32196491>.
- Teague, T. K., C. Tan, J. H. Marino, B. K. Davis, A. A. Taylor, R. W. Huey, and C. J. Van De Wiele (2010). “CD28 expression redefines thymocyte development during the pre-T to DP transition”. In: *Int Immunol* 22.5, pp. 387–97. URL: <https://www.ncbi.nlm.nih.gov/pubmed/20203098>.
- Telleria, E. L., J. B. Benoit, X. Zhao, A. F. Savage, S. Regmi, T. L. Alves e Silva, M. O’Neill, and S. Aksoy (2014). “Insights into the trypanosome-host interactions revealed through transcriptomic analysis of parasitized tsetse fly salivary glands”. In: *PLoS Negl Trop Dis* 8.4, e2649. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24763140>.
- Tetty, M. D., F. Rojas, and K. R. Matthews (2022). “Extracellular release of two peptidases dominates generation of the trypanosome quorum-sensing signal”. In: *Nat Commun* 13.1, p. 3322. URL: <https://www.ncbi.nlm.nih.gov/pubmed/35680928>.
- Tomlinson, S., F. Vandekerckhove, U. Frevert, and V. Nussenzweig (1995). “The induction of *Trypanosoma cruzi* trypomastigote to amastigote transformation by low pH”. In: *Parasitology* 110 (Pt 5), pp. 547–54. URL: <https://www.ncbi.nlm.nih.gov/pubmed/7541124>.
- Traag, V. A., L. Waltman, and N. J. van Eck (2019). “From Louvain to Leiden: guaranteeing well-connected communities”. In: *Sci Rep* 9.1, p. 5233. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30914743>.
- Tran, H. T. N., K. S. Ang, M. Chevrier, X. Zhang, N. Y. S. Lee, M. Goh, and J. Chen (2020). “A benchmark of batch-effect correction methods for single-cell RNA sequencing data”. In: *Genome Biol* 21.1, p. 12. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31948481>.
- Trapnell, C., D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn (2014). “The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells”. In: *Nat Biotechnol* 32.4, pp. 381–386. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24658644>.
- Tung, P. Y., J. D. Blischak, C. J. Hsiao, D. A. Knowles, J. E. Burnett, J. K. Pritchard, and Y. Gilad (2017). “Batch effects and the effective design of single-cell gene expression studies”. In: *Sci Rep* 7, p. 39921. URL: <https://www.ncbi.nlm.nih.gov/pubmed/28045081>.
- Urban, I., L. B. Santurio, A. Chidichimo, H. Yu, X. Chen, J. Mucci, F. Aguero, and C. A. Buscaglia (2011). “Molecular diversity of the *Trypanosoma cruzi* TcSMUG family of mucin genes and proteins”. In: *Biochem J* 438.2, pp. 303–13. URL: <https://www.ncbi.nlm.nih.gov/pubmed/21651499>.
- Urwyler, S., E. Studer, C. K. Renggli, and I. Roditi (2007). “A family of stage-specific alanine-rich proteins on the surface of epimastigote forms of *Trypanosoma brucei*”. In: *Mol Microbiol* 63.1, pp. 218–28. URL: <https://www.ncbi.nlm.nih.gov/pubmed/17229212>.
- Van Den Abbeele, J., Y. Claes, D. van Bockstaele, D. Le Ray, and M. Coosemans (1999). “*Trypanosoma brucei* spp. development in the tsetse fly: characterization of the post-mesocyclic stages in the foregut and proboscis”. In: *Parasitology* 118 (Pt 5), pp. 469–78. URL: <https://www.ncbi.nlm.nih.gov/pubmed/10363280>.
- Van den Berge, K., H. Roux de Bezieux, K. Street, W. Saelens, R. Cannoodt, Y. Saeys, S. Dudoit, and L. Clement (2020). “Trajectory-based differential expression analysis for single-cell sequencing data”. In: *Nat Commun* 11.1, p. 1201. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32139671>.
- Varki, A. (2011). “Since there are PAMPs and DAMPs, there must be SAMPs? Glycan ”self-associated molecular patterns” dampen innate immunity, but pathogens can mimic them”. In: *Glycobiology* 21.9, pp. 1121–4. URL: <https://www.ncbi.nlm.nih.gov/pubmed/21932452>.
- Vigneron, A., M. B. O’Neill, B. L. Weiss, A. F. Savage, O. C. Campbell, S. Kamhawi, J. G. Valenzuela, and S. Aksoy (2020). “Single-cell RNA sequencing of *Trypanosoma brucei* from tsetse salivary glands

- unveils metacyclogenesis and identifies potential transmission blocking antigens”. In: *Proc Natl Acad Sci U S A* 117.5, pp. 2613–2621. URL: <https://www.ncbi.nlm.nih.gov/pubmed/31964820>.
- Virshup, Isaac, Sergei Rybakov, Fabian J. Theis, Philipp Angerer, and F. Alexander Wolf (2021). “ann-data: Annotated data”. In: *The Journal of Open Source Software*.
- Voncken, F., F. Gao, C. Wadforth, M. Harley, and C. Colasante (2013). “The phosphoarginine energy-buffering system of trypanosoma brucei involves multiple arginine kinase isoforms with different subcellular locations”. In: *PLoS One* 8.6, e65908. URL: <https://www.ncbi.nlm.nih.gov/pubmed/23776565>.
- Wagner, D. E. and A. M. Klein (2020). “Lineage tracing meets single-cell omics: opportunities and challenges”. In: *Nat Rev Genet* 21.7, pp. 410–427. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32235876>.
- Wan, Y. Y. (2014). “GATA3: a master of many trades in immune regulation”. In: *Trends Immunol* 35.6, pp. 233–42. URL: <https://www.ncbi.nlm.nih.gov/pubmed/24786134>.
- Wang, S., A. O. Pisco, A. McGeever, M. Brbic, M. Zitnik, S. Darmanis, J. Leskovec, J. Karkanas, and R. B. Altman (2021). “Leveraging the Cell Ontology to classify unseen cell types”. In: *Nat Commun* 12.1, p. 5556. URL: <https://www.ncbi.nlm.nih.gov/pubmed/34548483>.
- Weelden, S. W. van, B. Fast, A. Vogt, P. van der Meer, J. Saas, J. J. van Hellemond, A. G. Tielens, and M. Boshart (2003). “Procyclic Trypanosoma brucei do not use Krebs cycle activity for energy generation”. In: *J Biol Chem* 278.15, pp. 12854–63. URL: <https://www.ncbi.nlm.nih.gov/pubmed/12562769>.
- Welch, C., M.K. Santra, W. El-Assaad, X. Zhu, W.E. Huber, R.A. Keys, J.G. Teodoro, and M.R. Green (2009). “Identification of a protein, G0S2, that lacks Bcl-2 homology domains and interacts with and antagonizes Bcl-2”. In: *Cancer Research* 69.17, pp. 6782–6789.
- Wendt, G., L. Zhao, R. Chen, C. Liu, A.J. O’Donoghue, C.R. Caffrey, M.L. Reese, and J.J. Collins (2020). “A single-cell RNA-seq atlas of Schistosoma mansoni identifies a key regulator of blood feeding”. In: *Science* 369, pp. 1644–1649.
- Williams, K. L., I. Nanda, G. E. Lyons, C. T. Kuo, M. Schmid, J. M. Leiden, M. H. Kaplan, and E. J. Taparowsky (2001). “Characterization of murine BATF: a negative regulator of activator protein-1 activity in the thymus”. In: *Eur J Immunol* 31.5, pp. 1620–7. URL: <https://www.ncbi.nlm.nih.gov/pubmed/11466704>.
- Wirtz, E., S. Leal, C. Ochatt, and G.A.M. Cross (1999). “A tightly regulated inducible expression system for conditional gene knock-outs and dominant-negative genetics in Trypanosoma brucei”. In: *Molecular and Biochemical Parasitology* 99, pp. 89–101.
- Wolf, F. A., P. Angerer, and F. J. Theis (2018). “SCANPY: large-scale single-cell gene expression data analysis”. In: *Genome Biol* 19.1, p. 15. URL: <https://www.ncbi.nlm.nih.gov/pubmed/29409532>.
- Wolf, F. A., F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Gottgens, N. Rajewsky, L. Simon, and F. J. Theis (2019). “PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells”. In: *Genome Biol* 20.1, p. 59. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30890159>.
- Xi, N. M. and J. J. Li (2021). “Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data”. In: *Cell Syst* 12.2, 176–194 e6. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33338399>.
- Xiao, H., K. Rasul, and R. Vollgraf (2017). “Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms”. In: *ArXiv*.
- Xue, Y., T. C. Theisen, S. Rastogi, A. Ferrel, S. R. Quake, and J. C. Boothroyd (2020). “A single-parasite transcriptional atlas of Toxoplasma Gondii reveals novel control of antigen expression”. In: *Elife* 9. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32065584>.
- Yang, S., S. E. Corbett, Y. Koga, Z. Wang, W. E. Johnson, M. Yajima, and J. D. Campbell (2020). “Decontamination of ambient RNA in single-cell RNA-seq with DecontX”. In: *Genome Biol* 21.1, p. 57. URL: <https://www.ncbi.nlm.nih.gov/pubmed/32138770>.

- Yoshida, N., S. Favoreto Jr, A.T. Ferreira, and P.M. Manque (2000). “Signal transduction induced in *Trypanosoma cruzi* metacyclic trypomastigotes during the invasion of mammalian cells”. In: *Brazilian Journal of Medical and Biological Research* 33, pp. 269–278.
- Young, M. D. and S. Behjati (2020). “SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data”. In: *Gigascience* 9.12. URL: <https://www.ncbi.nlm.nih.gov/pubmed/33367645>.
- Zamze, S. E., M. A. Ferguson, R. Collins, R. A. Dwek, and T. W. Rademacher (1988). “Characterization of the cross-reacting determinant (CRD) of the glycosyl-phosphatidylinositol membrane anchor of *Trypanosoma brucei* variant surface glycoprotein”. In: *Eur J Biochem* 176.3, pp. 527–34. URL: <https://www.ncbi.nlm.nih.gov/pubmed/2458923>.
- Zappia, L. and A. Oshlack (2018). “Clustering trees: a visualization for evaluating clusterings at multiple resolutions”. In: *Gigascience* 7.7. URL: <https://www.ncbi.nlm.nih.gov/pubmed/30010766>.
- Zhou, W., F. Gao, M. Romero-Wolf, S. Jo, and E.V. Rothenberg (2022). “Single-cell deletion analyses show control of pro-T cell developmental speed and pathway by Tcf7, Spi1, Gata3, Bcl11a, Erg, and Bcl11b”. In: *Science* 7.71.
- Ziegelbauer, K., M. Quinten, H. Schwarz, T. W. Pearson, and P. Overath (1990). “Synchronous differentiation of *Trypanosoma brucei* from bloodstream to procyclic forms in vitro”. In: *Eur J Biochem* 192.2, pp. 373–8. URL: <https://www.ncbi.nlm.nih.gov/pubmed/1698624>.