TrAGEDy: Trajectory Alignment of Gene Expression Dynamics

Ross F. Laidlaw[1], Emma M. Briggs[1,2], Keith R. Matthews[2], Richard McCulloch[1], Thomas D. Otto[1*]

[1]Wellcome Centre for Integrative Parasitology, School of Infection & Immunity, University of Glasgow, Glasgow, UK
[2]Institute for Immunology and Infection Research, School of Biological Sciences, University of Edinburgh, Edinburgh, UK

*Correspondence to thomasdan.otto@glasgow.ac.uk

Abstract

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. We 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.

Background

First described in 2014 (1), Trajectory Inference (TI) methods order cells based on the gradual change in transcript levels underlying biological process/es captured by single cell

26 RNA-sequencing (scRNA-seq) (2). Cells are assigned a “pseudotime” value based on their
27 position in the inferred trajectory, allowing differential gene expression tests over
28 pseudotime. Various methods have been developed to identify differentially expressed (DE)
29 genes in a biological process of interest, in some cases allowing targeted comparisons.
30 Notably, TradeSeq (3) uses negative binomial general additive models (GAM) and Wald
31 tests to assess whether a gene is differentially expressed over pseudotime within a trajectory
32 or between lineages of a trajectory. Alternatively, Monocle’s BEAM (4)identifies genes
33 associated with a particular branch of a trajectory whereas pseudotimeDE (5) uses
34 permutation to account for uncertainty in pseudotime and a zero-inflated negative binominal
35 GAM to account for expression value dropout. By applying these techniques, dynamic gene
36 expression patterns associated with a variety of interesting biological events have been
37 investigated, such as regulators of myogenesis (1), effector gradients in CD4+ T cells (6),
38 and life cycle transitions of the pathogen *Trypanosoma brucei* (7).

39 These methods can be extended to identify DE genes between two conditions where the
40 common biological process under analysis varies, such as before and after mutation or in an
41 overlapping response to different stimuli. One approach is to integrate the separate datasets
42 together and do a cluster-based comparison between conditions, treating cells as being at a
43 discrete stage in development captured in each cluster. This method misses the nuance that
44 development may be continuous, with some cells falling between stage boundaries. Another
45 approach is to integrate the datasets together, perform TI to capture the shared, possibly
46 branching, trajectory and differential expression tests across this pseudotime axis for each
47 condition, using tools like TradeSeq’s conditionTest. However, the integration may force
48 similarity between the trajectories, thus obscuring DE genes.
49

50 Alternatively, trajectory alignment can align cells from independently generated trajectories
51 to find a common pseudotime axis that retains the original ordering of cells without dataset
52 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 (8), 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 limitations on the type of alignment that can be analysed. Trajectories must start and end at the same point, 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 which may not be represented on the opposing trajectory, are nevertheless matched to a point on the other trajectory, making interpreting the subsequent alignment difficult. Here, we build on the work of cellAlign and develop a tool we called 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, allowing us to implement our own differential gene expression test, 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 *Trypanosoma brucei* lifecycle and T cell development under different genetic conditions, revealing gene expression changes undetected by current methods.

Results

TrAGEDy overview

A typical workflow for TrAGEDy is as follows (Fig.1). Prior to trajectory alignment with TrAGEDy, a predefined list of genes is selected for TI (Trajectory Inference) and dissimilarity is calculated between the datasets. Top variable genes or cluster markers can be selected, for example. The two datasets are then projected into a reduced dimension space using the preselected features and TI performed on the two datasets independently (Fig.1A). There is no limitation on what TI method can be used, provided 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 which controls a gaussian pseudotime window, the user can alter the contribution level more distal cells have to interpolated point gene expression (Fig.1B).

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. We used Spearman correlation, as it is less sensitive to outliers than Pearson and, unlike Euclidean distance, does not require scaling of gene expression values.

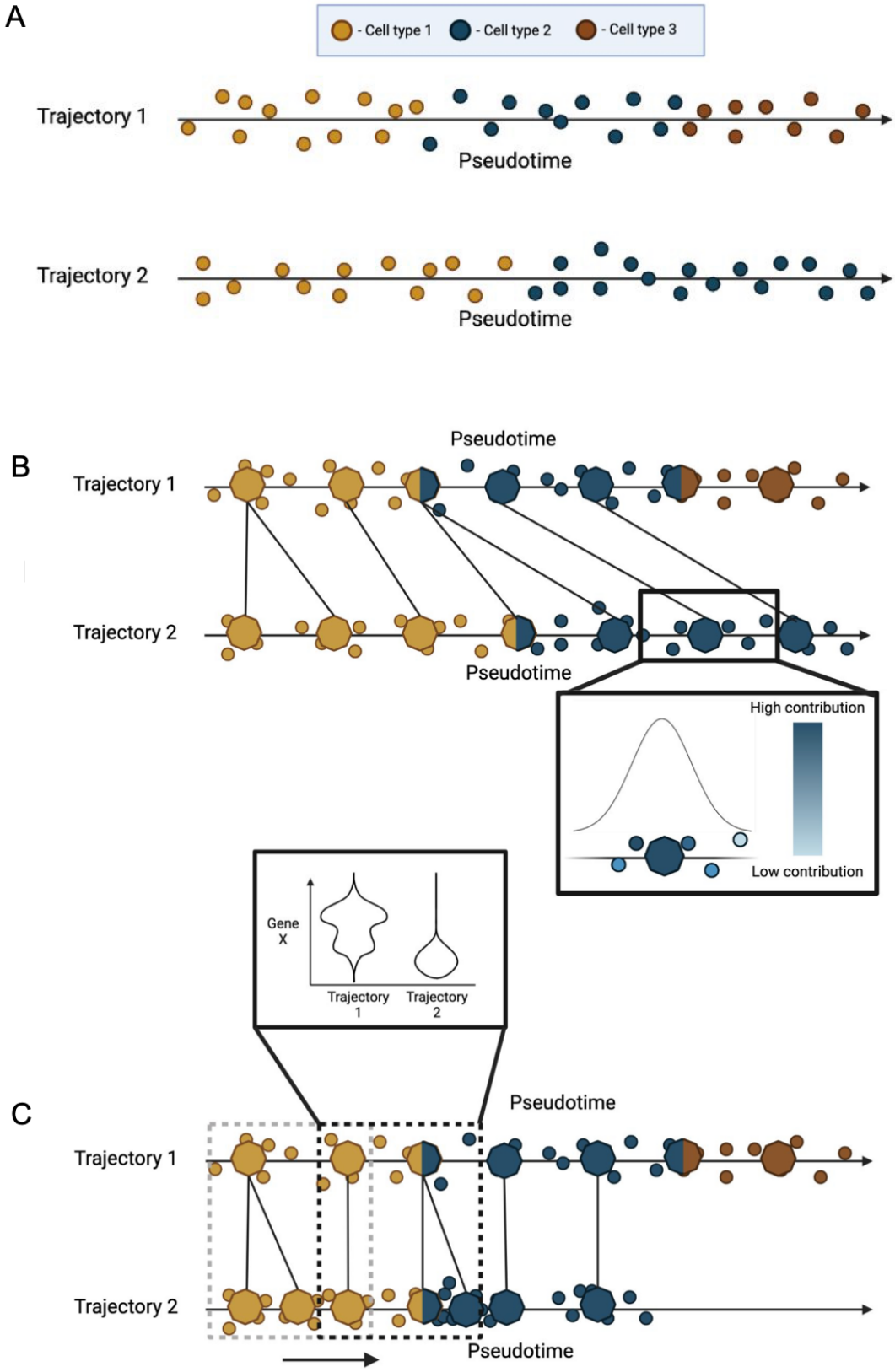


Figure 1. Graphical overview of the TrAGEDy process.

Trajectory Inference is carried out on two datasets which 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, to extract differentially expressed 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 (Fig.1B). 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 which are most transcriptionally similar to one another. 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 which may have diverged in the middle of their respective trajectories to be compared. 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.1C).

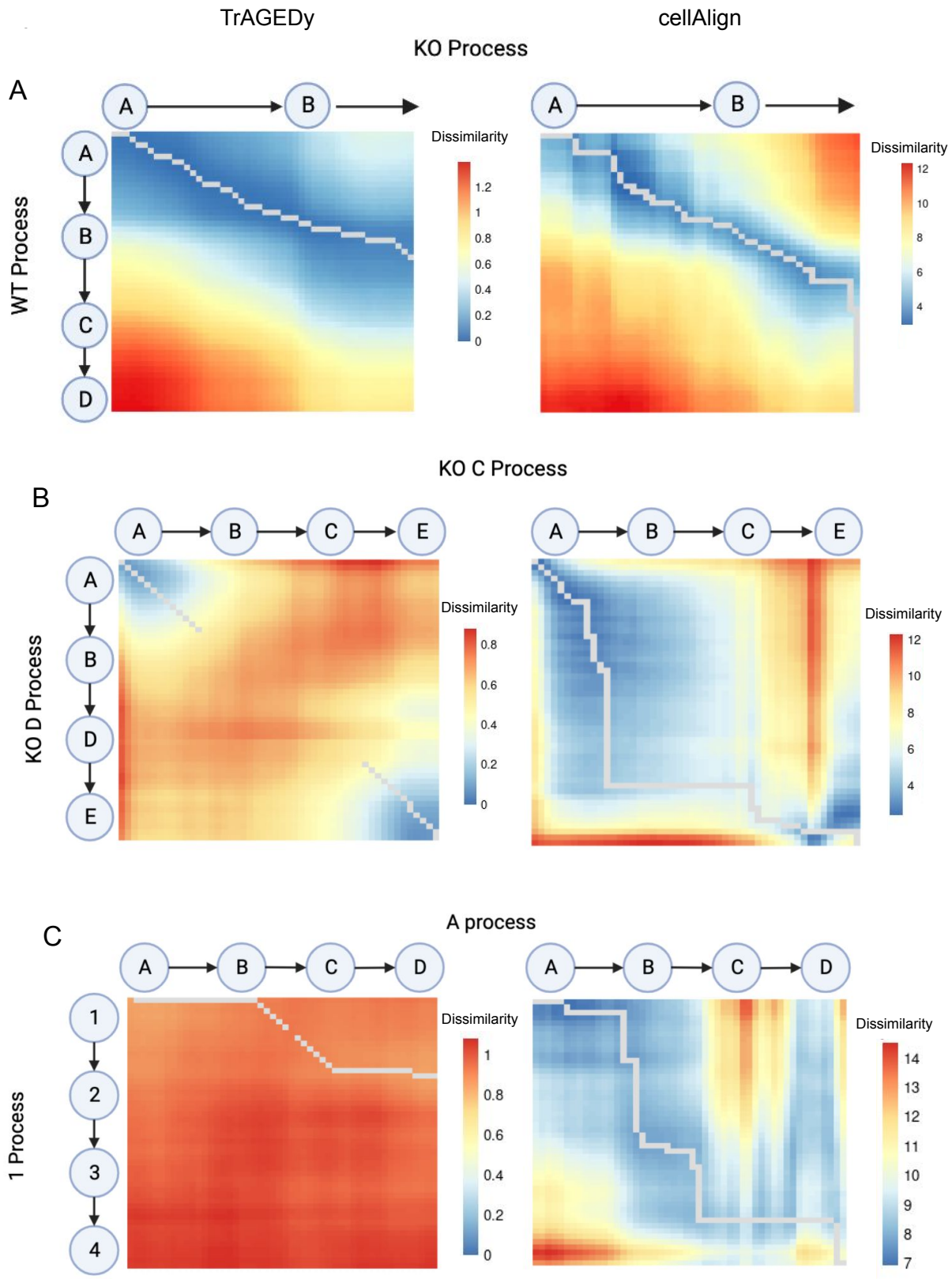
To extract DE genes, we performed a sliding window comparison, clustering cells at similar points in aligned pseudotime together then testing for significance with a Mann-Whitney U test, extracting genes that are DE at different points in the shared process (Fig.1C). Clustering with a sliding window allows comparisons over a continuous process. Users can define how much overlap there is between the windows of comparison.

Aligning simulated scRNA-seq datasets

We tested TrAGEDy's ability to correctly identify the alignment of two trajectories on datasets where the alignment points were known *a priori*. Using Dyngen (9) 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.

The first positive test case used two datasets, the first simulating a full developmental progression involved sequential expression of many genes (wild type; WT) and the second resulting from knockout (KO) of one of these genes (Fig.2A). Thus, the KO trajectory is truncated at the point of gene KO. 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, CellAlign was also performed using these datasets, with Euclidean distance selected as the metric to assess similarity, as this was the metric used in the source publication. Like TrAGEDy, CellAligns 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.


TrAGEDy
cellAlign
KO Process
A
WT Process
Dissimilarity
KO C Process
B
KO D Process
A process
C
1 Process

144

145

146 **Figure 2. Alignment of simulated single cell RNA sequencing datasets with TrAGEDy**
147 **and cellAlign.**
148 Three simulated datasets were generated with Dyngen in which there was initial alignment
149 (A), alignment at the beginning and end only (B) and no alignment (C), between the two
150 simulated processes. TrAGEDy (left column) and cellAlign (right column) were used to align
151 the datasets with Euclidean distance being used to measure transcriptional dissimilarity in
152 cellAlign, while TrAGEDy used Spearman correlation. The transcriptional dissimilarity score
153 between individual interpolated points of the two processes is shown in the corresponding
154 key, with blue being low dissimilarity and red being high. If the interpolated points are part of
155 the optimal path, their index colour is grey.
156
157 To assess if TrAGEDy can align trajectories that have shared start and end points, but
158 deviate in the middle, we simulated two datasets with Dyngen that have a diverge-converge
159 backbone. By knocking out one of the divergent branches in one dataset (KO C) and the
160 other divergent branch in the other dataset (KO D), we simulated linear trajectories which do
161 not share a common process in their middle points (Fig.2B). Applying TrAGEDy to this
162 scenario accurately captured the initial and terminal alignment of the two trajectories, while
163 leaving the middle sections unaligned. CellAlign failed to align KO C and KO D correctly, as
164 CellAlign does not include features to prune matches.
165
166 For the negative control, two datasets with distinct gene regulatory networks and
167 transcription kinetics were generated (A and one) (Fig.2C). As there is no shared process
168 between the conditions, the dissimilarity scores are expected to be high, and no alignment
169 should be found. The dissimilarity score calculated by TrAGEDy between process A and
170 process one showed that most of the indexes have a dissimilarity score close to 1, and thus
171 do not share a common process. As TrAGEDy finds the optimal path by looking at the
172 context of the whole dissimilarity matrix, even if the dissimilarity scores are high, it will still
173 find a path, as seen in the negative control. TrAGEDy will give a warning to the user if the

median of the cut dissimilarity matrix scores is over 0.6 and thus it is up to the user 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 ~7-14, while the other two datasets ranged from ~4-12. If, instead, Pearson correlation is selected for CellAlign, it becomes clearer that the two datasets share little transcriptional similarity (supp. Fig.4C).

WT vs Bcl11b KO T cell development analysis

Simulated data represents a ‘best case’ scenario for benchmarking tools, as they often lack the complexity of real datasets. We thus decided to apply TrAGEDy to real scRNA-seq datasets where the underlying process has been characterised. We first applied TrAGEDy to a scRNA-seq dataset of *in vitro* T cell development in WT and Bcl11b KO condition mice over two timepoints (10). The data contain cells undergoing the early stages of T cell development in the thymus from thymus-seeding precursors (TSP) to the double negative three stage (DN3), where the cells undergo beta selection (11). The authors find that when Bcl11b is removed, the cells deviate from the T cell commitment pathway and develop distinct transcriptomic signatures, including expression of Id2 and Rora. However, some Bcl11b KO cells still express markers of T cell pathway commitment such as the gene for the pre-T cell Antigen receptor alpha (Ptcra) (12) and Rag1, one of the central genes for TCR gene (13) (supp. Fig. 6). We thus aimed to compare the WT and Bcl11b KO trajectories as development proceeds and diverges between the two conditions to profile the differences in transcriptome.

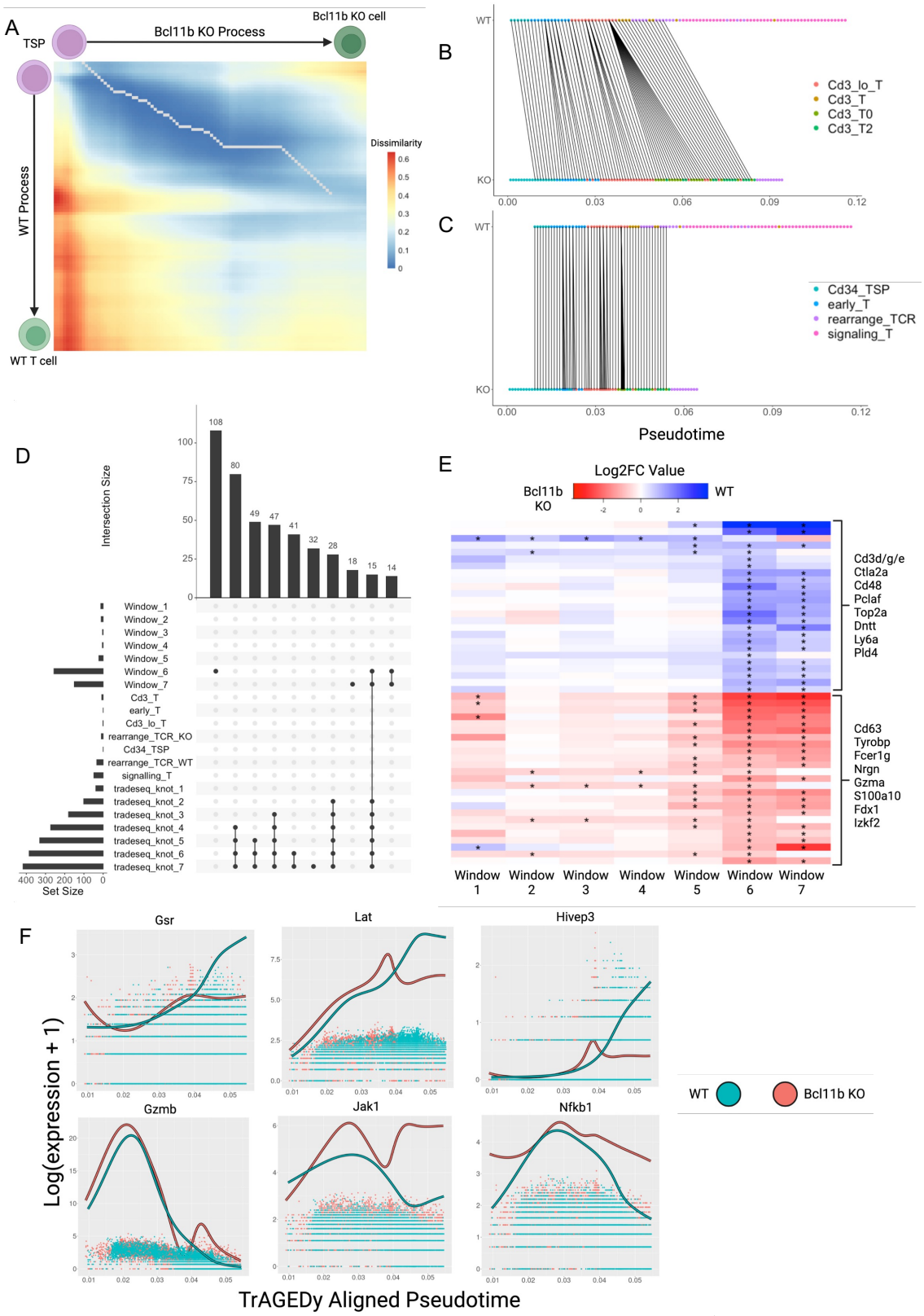


Figure 3. Analysis of WT and Bcl11b KO trajectories of T cell development with TrAGEDy

200 TrAGEDy alignment of the WT and Bcl11b KO trajectories of T cell development with
201 transcriptional dissimilarity calculated with Spearman correlation (A). Alignment of the
202 interpolated points of the WT and Bcl11b KO trajectories, showing connections between the
203 two processes as calculated by TrAGEDy. The interpolated points are displayed across
204 pseudotime before (B) and after (C) the pseudotime values have been modified by
205 TrAGEDy. (D) UpSet plot showing the top 10 intersections of differentially expressed genes
206 captured by TrAGEDy across 7 windows of comparison and Seurat over 7 clusters and
207 TradeSeq over 7 knot comparisons. (E) Heatmap showing DE genes with the highest log 2
208 fold-change (Log2FC) captured (but not uniquely captured) between the WT and Bcl11b KO
209 trajectories at 7 different windows of comparison. Each index is coloured by the Log2FC (red
210 = higher expression in KO, blue = higher expression in WT) and an * symbol indicates the
211 difference is significant (bonferroni adjusted $p < 0.05$, Mann-Whitney U test, abs(Log2FC) >
212 0.5, minimum percentage expressed > 0.1). Full list of heatmap genes can be found in
213 supplementary table 7. (F) TradeSeq plotSmoother plots showing the expression changes
214 that occur across the WT (blue) and Bcl11b KO (red) trajectories within the cells of aligned
215 pseudotime, as calculated by TrAGEDy, which share a common process. The 6 genes
216 plotted were only identified as being significantly DE by TrAGEDy.
217
218 We first tested whether TrAGEDy could overcome batch effects between the two replicates
219 to still return a biologically accurate alignment. The cells were sequenced over two runs, but
220 each run contained a variety of biological and technical replicates, as well as cell from
221 different sampling days. For simplicity, the WT cells sequenced in run 1 will be referred to as
222 WT1, the Bcl11b KO cells sequenced in run 1 will be referred to as Bcl11b KO1 and those
223 sequenced in run 2 as WT2 and Bcl11b KO2, resulting in four individual datasets. PHATE
224 (14) embeddings were generated independently and used as the basis for Slingshot TI to
225 calculate pseudotime values for each of the cells. When projected into the PHATE space,
226 some of the WT1 cells were separated from the main body of the trajectory, which had
227 downstream effects on the alignment, and so these cells were removed as they did not

228 appear to constitute part of the biological process under examination (supp. Fig. 7).
229 TrAGEDy was used to first align the two datasets for each condition together, resulting in
230 two separate alignments: a WT alignment of WT1 and WT2, and a Bcl11b KO alignment of
231 Bcl11b KO1 and Bcl11b KO2 (supp. Fig. 8). The WT1 dataset only contains cells from day
232 10 of sampling while WT2 only contains cells from day 13, meaning the alignment may be
233 skewed with the WT1 process starting before the WT2 process, but finishing before the WT2
234 process; a pattern captured by TrAGEDy (supp. Fig. 8). TrAGEDy was then performed on
235 the aligned WT and Bcl11b KO trajectories. The resulting final alignment shows the two
236 conditions sharing an initial common process, but the Bcl11b KO developmental progression
237 finished before that of the WT (Fig. 3A, B, C).

	TrajDE	tradeSeq	Seurat
----------0.25 Log2FC threshold----------			
Total genes	1064	980	196
Unique genes	554	479	12
----------0.5 Log2FC threshold----------			
Total genes	300	484	40
Unique genes	164	348	5

Runtime (minutes)	2.7	450	0.13

238

239 **Table 1 Summary of DE gene numbers for the WT vs Bcl11b KO T cell comparison.**
240 For the WT vs Bcl11b KO T cell development dataset comparison, the number of DE genes
241 captured by the three different methods in total, and DE genes that they captured uniquely
242 by each method, using two different Log2FC cutoffs (0.25 (top) and 0.5 (bottom)) for
243 comparisons. The time taken to run the necessary functions to get differentially expressed
244 genes (in minutes) is shown on the final row.

245

246 By aligning trajectories rather than integrating the datasets, we expected to preserve more
247 variability between them, which may be reflected in the DE genes captured. We thus

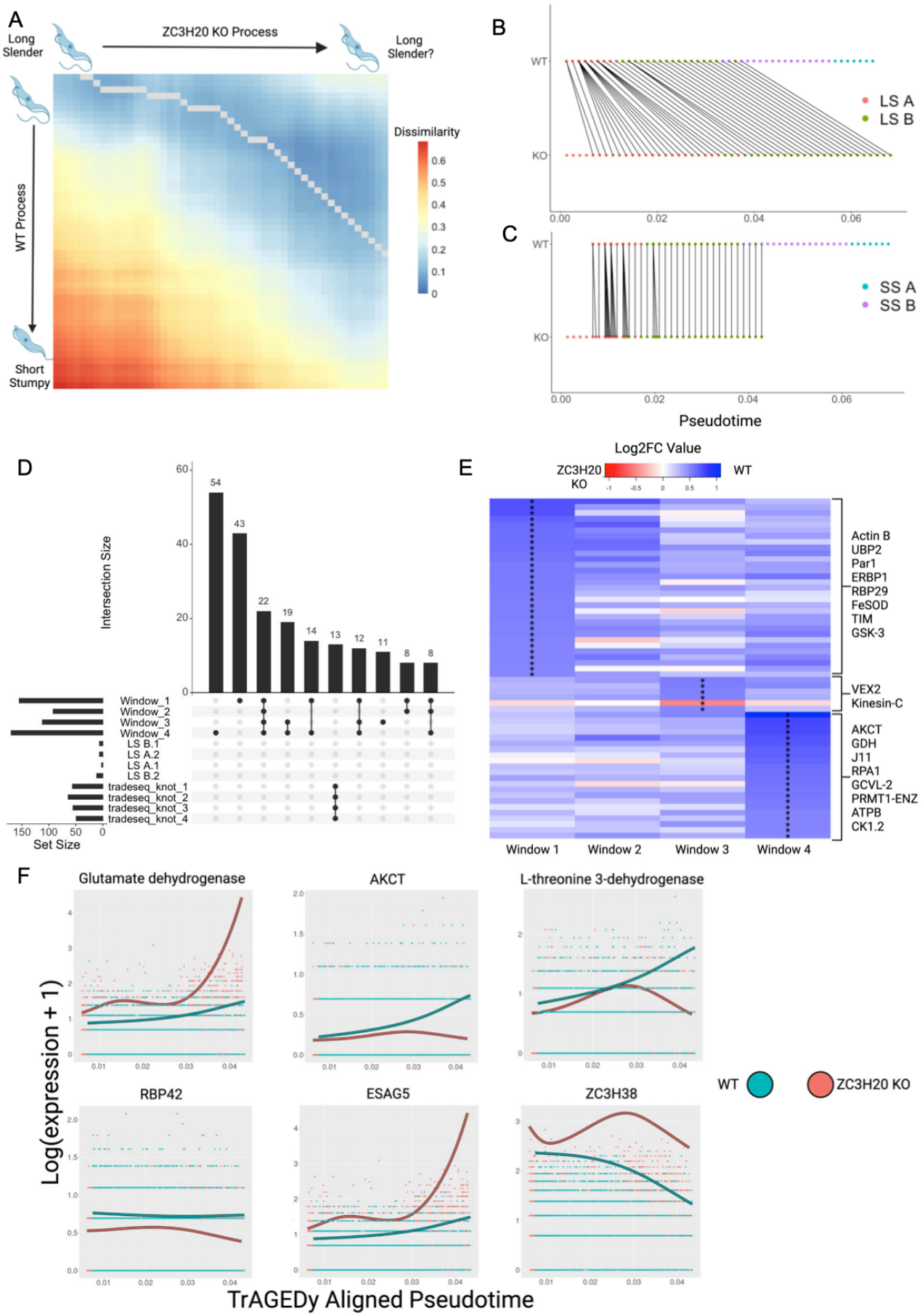
compared TrAGEDy against other methods which can extract DE genes between conditions. We compared the differential gene expression tests by TrAGEDy (TrajDE) with established methods of extracting DE genes: we compared TrajDE to a standard Seurat pipeline of comparing DE genes between conditions, and TradeSeq's method of comparing DE genes across trajectories from different conditions. Analysis was carried out using each method, and genes were classified as DE using two different thresholds, 0.25 and 0.5 log fold change (log2FC). A brief overview of these results can be seen in Table 1, which indicates that TrajDE was able to detect the most DE genes, including 554 not detected by TradeSeq or Seurat, at the lower threshold of 0.25 log2FC. Below we discuss biological implications of the genes classified as DE with the more stringent threshold of 0.5 log2FC. With this threshold, TrajDE returned 300 genes as being DE across the 7 windows of comparison, while Seurat only returned 40 across 7 clusters, and TradeSeq returned 484 genes across 7 comparison knots. TrajDE uniquely captured 164 genes that neither Seurat nor TradeSeq captured, whereas Seurat found 5 unique DE genes and TradeSeq uniquely captured 348 DE genes (table 1). The number of DE genes that were captured per window/cluster and the intersection of the different lists can be viewed in Fig.3C. Of the 164 DE genes captured only by TrAGEDy, 121 were expressed higher in the WT condition and 45 were expressed higher in the Bcl11b KO condition. Genes with higher expression in the WT include those associated with T cell signaling (Lat (12)), cell cycle associated genes (Top2a, Dut, Cenpa (15), Btg1 (16)), and those with functions for protecting against reactive oxygen species (ROS) (Txn1, Prdx1, Gsr) (17). The genes only TrAGEDy identified as being significantly higher expressed in Bcl11b KO cells include factors involved in immune cell activation (Jak1, Nfkb1 (18), Jund (12)), cytotoxic elements (Srgn, (19)), and transcription factors (Runx3 (20), Sox4 (21)).

Both TrAGEDy and TradeSeq captured many of the same biological relevant genes. Genes TrAGEDy captured as being significantly higher expressed in the WT, which TradeSeq also captured as being DE, included the pre-TCR alpha chain gene (Ptcra), CD3 chain genes

276 (Cd3e, Cd3g and Cd3d) (12), TCR recombination associate gene Dntt (22), immune
277 synapse associated genes (Cd48 (23), Cd7 (24,25)). Genes identified as significantly higher
278 expressed in the Bcl11b KO T cells by both methods included: NK cell marker (Cd244) (26),
279 cytotoxic molecule Gzma and IgE receptor Fcer1g (27). While it took TradeSeq around 7
280 hours to return its DE results (time taken to execute the fitGAM and conditionTest functions),
281 it only took TrAGEDy around 3 minutes to perform its process (Table 1). TrAGEDy
282 additionally allows the relative timing of expression changes between conditions across the
283 trajectory to be identified, by investigation of expression at each window (Fig. 3E). Plotting
284 some of the DE genes that only TrAGEDy captured show they are concurrent with the
285 patterns of expression TrAGEDy suggests, with Notch1, Rag1 and Hivep3 being higher
286 expressed in WT T cells, while Jak1, Gzmb and Nfkb1 are higher expressed in Bcl11b KO
287 cells (Fig. 3F).
288

289 WT vs ZC3H20 KO *Trypanosoma brucei* alignment

290 In its mammalian host, WT *Trypanosoma brucei* parasites transition from replicative slender
291 forms to non-replicative stumpy forms, through a quorum sensing mechanism (28–30)
292 Mutant *T. brucei* where an RNA binding protein essential for slender to stumpy
293 differentiation, ZC3H20, has been knocked out (ZC3H20 KO) are unable to undergo this
294 transition (31,32) and fail to express stumpy associated genes. Briggs and colleagues (7)
295 used 10x Chromium to sequence two biological replicates of WT *T. brucei* parasites
296 undergoing this transition *in vitro*, as well as ZC3H20 KO failing to differentiate in the same
297 conditions. The authors defined four main clusters Long Slender (LS) A and B, and Short
298 Stumpy (SS) A and B to show the progression of the transition, with LS A being the start and
299 SS B being the end. Thus, the final alignment is expected to show that the WT and the
300 ZC3H20 KO datasets share an initial common process, but that the KO process is truncated
301 or branched relative to the WT, as suggested by previous analysis using data integration and
302 Slingshot TI.


A
Long Slender
ZC3H20 KO Process
Long Slender?
WT Process
Short Stumpy
Dissimilarity
0.6
0.5
0.4
0.3
0.2
0.1
0
B
WT
KO
LS A
LS B
0.00
0.02
0.04
0.06
C
WT
KO
SS A
SS B
0.00
0.02
0.04
0.06
Pseudotime
D
Intersection Size
60
40
20
0
54
43
22
19
14
13
12
11
8
8
Window_1
Window_2
Window_3
Window_4
LS B.1
LS A.2
LS A.1
LS B.2
tradeseq_knot_1
tradeseq_knot_2
tradeseq_knot_3
tradeseq_knot_4
150 100 50 0
Set Size
E
Log2FC Value
ZC3H20 KO
WT
-1 -0.5 0 0.5 1
Actin B
UBP2
Par1
ERBP1
RBP29
FeSOD
TIM
GSK-3
VEX2
Kinesin-C
AKCT
GDH
J11
RPA1
GCVL-2
PRMT1-ENZ
ATPB
CK1.2
Window 1
Window 2
Window 3
Window 4
F
Glutamate dehydrogenase
AKCT
L-threonine 3-dehydrogenase
RBP42
ESAG5
ZC3H38
Log(expression + 1)
WT
ZC3H20 KO
TrAGEDy Aligned Pseudotime

303

304

305 **Figure 4. Analysis of WT and ZC3H20 KO *Trypanosoma brucei* development with**

306 **TrAGEDy.**

307 TrAGEDy alignment of the WT and ZC3H20 KO trajectories of *Trypanosoma brucei*

308 development with transcriptional dissimilarity calculated with Spearman correlation (A).

309 Alignment of the interpolated points of the WT and ZC3H20 KO trajectories, showing

310 connections between the two processes as calculated by TrAGEDy. The interpolated points

311 are displayed across pseudotime before (B) and after (C) the pseudotime values have been

312 modified by TrAGEDy. (D) UpSet plot showing the top 10 intersections of DE genes as

313 captured by TrAGEDy across four windows of comparison and Seurat over the four long

314 slender (LS) clusters and TradeSeq over 4 knot comparisons. (E) Heatmap showing the log

315 2 fold-change (log2FC) of genes found only by TrAGEDy to be DE between the WT and

316 ZC3H20 KO trajectories in only 1 of the 4 different windows of comparison. Each index is

317 coloured by the Log2FC (red = higher expression in KO, blue = higher expression in WT)

318 and a * symbol indicates the difference is significant (bonferroni adjusted $p < 0.05$ Mann-

319 Whitney U test, abs(Log2FC) > 0.5, minimum percentage expressed > 0.1). Genes names

320 were taken from TriTrypDB. Full list of heatmap genes can be found in Supplementary table

321 7. (F) TradeSeq plotSmoother plots showing the expression changes that occur across the

322 WT (blue) and ZC3H20 KO (red) trajectories within the cells of aligned pseudotime, as

323 calculated by TrAGEDy, which share a common process. The 6 genes plotted were only

324 identified as being significantly DE by TrAGEDy.

325

326

327 Applying TrAGEDy to WT01 and WT02 returned an alignment which started at the same

328 point, and ended just before the final endpoint, returning a nearly 1:1 alignment, as predicted

329 of biological replicates (supp. Fig.5). We next used the aligned WT values to perform

330 TrAGEDy alignment between the combined WT replicates and the ZC3H20 KO dataset. The

331 resulting alignment showed that the two process have an initial point of alignment with the

ZC3H20 KO process finishing before the WT process does (Fig.4A). Using the DTW calculated path, we could then adjust the pseudotime values of the trajectories to get a common pseudotime axis (Fig.4B, C).

	TrajDE	tradeSeq	Seurat
----------0.25 Log2FC threshold----------			
Total genes	604	253	179
Unique genes	344	36	10
----------0.5 Log2FC threshold----------			
Total genes	264	78	14
Unique genes	208	25	0

Runtime (minutes)	0.85	96	0.024

Table 2 Summary of DE gene numbers for the WT vs ZC3H20 KO *Trypanosoma brucei* comparison.

For the WT vs ZC3H20 KO *Trypanosoma brucei* development dataset comparison, the number of DE genes captured by the three different methods in total, and DE genes that they captured uniquely by each method, using two different Log2FC cutoffs (0.25 (top) and 0.5 (bottom)) for comparisons. The time taken to run the necessary functions to get differentially expressed genes (in minutes) is shown on the final row.

In this case, comparing DE results between methods revealed TrAGEDy was able to detect more genes compared to TradeSeq and Seurat, at both a lower and high log2FC threshold, and detect a greater number of DE genes missed by the other methods. Using a 0.5 log2FC threshold, TrajDE extracted 264 DE genes from 4 windows of comparison, Seurat extracted 14 across the 4 annotated slender clusters, and TradeSeq captured 78 genes across the 4 knot comparisons (Table 2). Of the genes only captured by TrajDE, most had significantly

351 higher expression in the WT (n=202) compared with the ZC3H20 KO (n=7) (Fig 4D). The
352 202 genes upregulated in the WT condition uniquely identified by TrAGEDy includes genes
353 involved in metabolism and catabolism of amino acids (glutamate dehydrogenase (33), L-
354 threonine 3-dehydrogenase, and 2-amino-3-ketobutyrate coenzyme A ligase (AKCT) (34)
355 and that act in the cytoskeleton and motor activity (actin-B and cytoskeleton-associated
356 proteins, kinesin-C and dynein (35)) and post-transcriptional regulation of gene expression
357 (Splicing factor TSR1 (36), RBP42 (37), UBP1 & UBP2 (38)). TrAGEDy was the only method
358 able to detect the known stumpy marker, Protein Associated with Differentiation (PAD)2 (39),
359 as being DE and upregulated in the WT trajectory, as expected. The 7 genes TrAGEDy
360 identified as being significantly higher expressed in the ZC3H20 KO cells include GAPDH,
361 expression site-associated gene 5 (ESAG5), and ZC3H38. By assessing DE over sequential
362 windows, genes with transient transcript level changes between conditions across
363 pseudotime could be identified (Fig. 4E).
364
365 To visualise changes in expression between conditions, we used TradeSeq to fit a GAM to
366 each gene in the WT and ZC3H20 KO datasets using the new TrAGEDy calculated
367 pseudotime to plot the smoothed expression. While the GAM was fitted across the whole
368 TrAGEDY aligned pseudotime, visualisations were made for only cells TrAGEDy defined as
369 being aligned between the two datasets and showed that the GAMs follow a similar pattern
370 as identified by TrAGEDy, with AKCT, L-threonine dehydrogenase, glutamate
371 dehydrogenase & RBP48 being higher expressed in the WT cells and ESAG5 and ZC3H38
372 being higher expressed in the ZC3H20 KO cells (Fig.4F).

373 Discussion

374 In this paper we present TrAGEDy, a tool for aligning and comparing single cell trajectories
375 between conditions. TrAGEDy allows the alignment of processes which may start and end at
376 different points by identifying the point at which dissimilarity increases from the minimum

377 dissimilarity point. TrAGEDy assesses alignments to remove matches where the dissimilarity
378 score is unaligned in portions of the trajectories, allowing TrAGEDy to additionally capture
379 alignments where the processes diverge and subsequently converge. TrajDE then identifies
380 differences in gene expression over pseudotime between the conditions by taking a sliding
381 window soft clustering approach, allowing it to identify transient changes in gene expression
382 over the shared process.
383
384 Across multiple simulated datasets, TrAGEDy can find the underlying alignment between
385 two processes, where cellAlign fails, due to its ability to prune matches and identify optimal
386 start and end points in the dissimilarity matrix (Figure 2A). Across biological replicates and
387 different conditions of real datasets, TrAGEDy overcomes batch effects to deliver accurate
388 alignments. We additionally find that applying TrAGEDy to real scRNA-seq datasets
389 identifies biologically relevant genes where TradeSeq and Seurat DE fail, with these
390 comparisons being performed in a fraction of the time when compared to TradeSeq (Table 1
391 & 2).
392
393 A key consideration when applying TrAGEDy is the density of cells across pseudotime.
394 However, the window TrAGEDy uses to calculate expression at the interpolated points can
395 be adjusted to account for regions with few cells. Notably, if cells are present which are
396 separated from the main bulk of the process, like in the WT1 T cell dataset (supp. Fig.7), this
397 leads to an uneven dissimilarity score between neighboring interpolated points, thus biasing
398 the end alignment. These outlier cells should thus be removed before TrAGEDy is carried
399 out, to ensure a smooth and accurate alignment.
400
401 As of 2019, there were more than 70 TI methods, all of which have varying performance,
402 functionality and constraints (2). Additionally, the chosen method of dimensionality reduction
403 (PCA, UMAP, PHATE, etc) used to generate the embeddings that the trajectory is calculated
404 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 (14). User discretion is advised when choosing the TI and dimensionality reduction method. Given this 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. In all simulated datasets discussed here, TrAGEDy was able to correctly align trajectories, and with greater accuracy than cellAlign. 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 conditions which is not implemented in most tools designed to specifically analyse trajectories, including pseudotimeDE (5) and Monocle (4). Low detection of DE genes by Seurat is thus likely explained by the fact it does not take pseudotime into account when carrying out the differential expression tests. TradeSeq, however, is designed for pseudotime-based differential gene expression tests and is able to compare conditions. We thus compared TrAGEDy results with these two tools, using both a low and higher stringency logFC threshold.

When analysing the T cell development dataset, TradeSeq captured more DE genes at a log2FC threshold of 0.5 than TrAGEDy, but at a log2FC of 0.25 TrAGEDy captured more DE genes. 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 beta-selection, before they begin to rearrange their TCR alpha chain, T cells undergo a proliferative burst and then become quiescent (40). In accordance with this, cell cycle genes are mostly DE in the final

433 two windows (windows 6 and 7) and the anti-proliferation factor Btg1 (16) (which was only
434 detected by TrAGEDy) is significantly higher expressed in the WT in only the final window
435 (window 7) (supp. Table 1). This indicates that TrAGEDy has managed to capture the
436 transient changes in proliferative state as WT T cells prepare to rearrange their TCR alpha
437 chain. It also indicates that Bcl11b KO cells may have reduced ability to carry out this
438 proliferative burst. TrAGEDy also captured the transcription factor Hivep3 as significantly
439 more highly expressed in the WT compared with Bcl11b KO cells. This gene has been
440 shown to bind regions of the DNA associated with V(D)J recombination of antigen receptors
441 in T and B cells (41). Interestingly, Dntt is also significantly higher expressed in the WT, and
442 so Bcl11b KO cells may have reduced capacity to rearrange their TCR; affecting their
443 signaling ability with a knock-on effect on the proliferative burst.
444
445 While the effect of ROS on T cell development has not been fully elucidated, increased ROS
446 generation leads to reduced levels of mature single positive T cells in the thymus and
447 treatment of T cells with factors which limit ROS show increased survival (42). This suggests
448 that reduced ROS positively influences T cell development, and with genes associated with
449 antioxidation being significantly higher expressed in WT cells, Bcl11b KO cells may be more
450 susceptible to ROS, limiting their maturation ability.
451
452 In the Bcl11b KO T cells, TrAGEDy identified Gzmb and Srgn as being significantly more
453 highly expressed compared with WT T cells. Gzmb is a key enzyme expressed in cytotoxic
454 immune cells such as Natural killer (NK) cells while Srgn plays a role in storage of Gzmb in T
455 cells (19). Expression of such genes (and Cd244 which both TradeSeq and TrAGEDy
456 detected) is in keeping with Zhou and colleagues' (10) observation that the Bcl11b KO T
457 cells appear to develop a NK cell-like phenotype.
458
459 While TradeSeq returns more unique DE genes than the other methods tested here for the T
460 cell dataset, it does so at the cost of runtime, which stretched into several hours, compared

with the minutes taken for TrAGEDy to run its process. As TrAGEDy not only uniquely captures biological relevant DE genes but captures many that TradeSeq does, at the fraction of the time, TrAGEDy is well poised to quickly and efficiently uncover the biological differences that occur across a process between conditions.

When applied to *T. brucei* development, TrAGEDy captured more DE genes compared to Seurat and TradeSeq at both low and high logFC cut-offs, suggesting greater analytical power. Many of the 208 genes identified only by TrAGEDy have known associations with the biological process under investigation (discussed below), indicating TrAGDEy can uncover DE genes with biological relevance as opposed to increasing the false-discovery rate.

For example, several DE genes reflect the parasite's need to alter its metabolism when moving from a high to low glucose environment. Glutamate hydrogenase, found to be significantly more highly expressed in the WT, is important in proline metabolism, a key feature of survival in the tsetse fly, as proline is a major energy source for the parasite (33). Stumpy form *T. brucei,* a specialised transmission stage, preadapts for survival in the tsetse fly midgut where they differentiate to procyclic forms, and so upregulation of glutamate hydrogenase is likely to aid rapid adjustment to nutrient availability when ingested during a bloodmeal (43). Other observations include WT upregulation of AKCT and L-threonine 3-dehydrogenase, both of which are involved in the catabolic pathway converting L-threonine into glycine (34). L-threonine is readily metabolised by procyclic forms and preferably metabolised over other amino acids in glucose-depleted environments (44), where it is used in lipid biosynthesis (45). Further reinforcing this metabolic change is TrAGEDy identifying RBP42 as being significantly more highly expressed in the WT. This protein has been implicated as a key regulator of metabolism associated genes (46), and it is significantly DE in the final window along with the previously mentioned genes. Thus, TrAGEDy captured the WT parasite pre-adaptation to the tsetse fly environment through the upregulation of genes that will help the parasite respond to oxidative stress and alter metabolism, and failure of the

489 ZC3H20 KO parasites to do the same. Taken together, TrAGEDy captures DE genes not
490 captured by other differential expression test methods, revealing greater insight into
491 functions that differ between the WT and ZC3H20 KO *T. brucei* developmental programmes.
492
493 TI represents an approximation of how a developmental trajectory might look and may not
494 represent the actual ordering of the cells through the process. Adding extra temporal
495 information from lineage tracing will help add more biological validity to single cell data (47).
496 TrAGEDy could thus take lineage trace inferred time, rather than pseudotime, as it's input to
497 allow the analysis of such datasets, allowing it to remain relevant as more sophisticated
498 single cell sequencing techniques are introduced. Furthermore, the advent of perturb-seq
499 allows researchers to more easily create and analyse the effect of genetic perturbations on
500 processes (48), creating multiple and diverse opportunities for TrAGEDy to be applied.
501
502 Both TrAGEDy and cellAlign are restricted to aligning linear processes due to DTW only
503 being able to operate with linear processes. To overcome this limitation, branching
504 processes could be treated as multiple root-to-leaf linear processes and aligned and
505 analysed in a pairwise fashion but how accurate this process would be is uncertain as the
506 two paths are not independent of one another. Thus, future work needs to be carried out to
507 identify a method permitting the alignment of complex trajectories in order to better expand
508 the datasets which trajectory alignment can be applied to.
509

510 Conclusion

511 Our paper demonstrates the ability of TrAGEDy to overcome batch effects between datasets
512 and reveal the underlying alignment between two datasets, identifying shared and dissimilar
513 points across the independent trajectories. For some cases, TrAGEDy can identify more DE
514 genes than other methods, and in all cases, TrAGEDy can identify biologically relevant DE

genes between conditions and at specific points in the trajectory. With the advent of perturb-seq methods, the need for TrAGEDy becomes 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 with scRNA-seq to reveal novel detail.

Methods

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 supplementary figure 1 and implementation of the method can be seen, in full, in the code contained on our GitHub.

Create interpolated points

Performing alignment directly on every cell in a dataset would not only be computationally and time expensive, but also prone to noise. As such, 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.

In the cellAlign method, each interpolated point is given a 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. In cellAlign, the size of the window is the same for every interpolated point. For TrAGEDy, the user specifies 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.

Scoring dissimilarity

For every interpolated point in the two trajectories, we calculate the transcriptome 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. Hence, all the dissimilarity scales start from 0 (i.e. least dissimilarity), and we change the correlation scores to be 1 minus the correlation score.

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 (supp. 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 (supp. 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 becomes the new start/end point. This repeats until all the points that come before/after it, have a dissimilarity score higher than the threshold (supp. Fig.2B). The matches that occur

before the new start point and after the new end point are removed and DTW is performed on this cut dissimilarity matrix, assigning indexes to either be matched (part of the shared process) or unmatched.

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. 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 (supp. Fig.2C).

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. As we have two processes, we have two vectors: $y_i^1\{i \in 1{:}m\}$ and, $y_j^2\{j \in 1{:}n\}$ 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 ith interpolated point in process 1 and y_j^2 being the pseudotime of the jth interpolated point in process 2.

Interpolated points may only match one another (supp. Fig.2A) 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 (supp. Fig.3B). For adjusting pseudotime of the former scenario, given that y_i^1 and y_j^2 are matched and that $y_j^2 < y_i^1$: $y_{i:m}^1 = y_{i:m}^1 + \left(y_j^2 - y_i^1\right)$ (supp. Fig.3A). For a multi match, first TrAGEDy aligns the pseudotime of the next match not involved in the current multi match (supp. Fig.3B, equation 1), using the steps outlined for the individual match, then it does the same for the first match in the multi match (supp. Fig.3B, equation 2). 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 (supp. Fig.3B, equation 3). This leaves us with an aligned pseudotime axis, where the matched interpolated points have similar pseudotime values (supp. Fig.3C). 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.

Differential expression analysis with TrAGEDy

TrAGEDy identifies DE genes across pseudotime, between the two conditions by taking a sliding window soft clustering approach.

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. To adjust the p-values to account for multiple testing, Bonferroni correction is used across all the genes in the dataset.

Simulating trajectories with Dyngen

All simulated datasets were simulated with Dyngen (9). 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 B2 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 GitHub link in the "Availability of data and materials" section.

Applying TrAGEDy 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 15 PCA dimensions clustering was performed at a resolution of 0.5 for all the datasets as well as UMAP. Slingshot (49) 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 Log2FC 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. Specific parameters such as number of dimensions used for each dataset can be found on the GitHub linked in the "Availability of data and materials" section.

Pre-processing of WT vs ZC3H20 KO *Trypanosoma brucei* dataset

The processed RDS (R dataset) file containing the final cluster labels was acquired from the authors (https://zenodo.org/record/5163554#.Yyb113bMKUI). 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’.

TrAGEDy analysis of WT vs ZC3H20 KO *Trypanosoma brucei* dataset

The integrated object was split into the individual datasets (WT01, WT02 and ZC3H20_KO). The top 300 DE genes 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 Log2FC was more than 0.5, bonferroni corrected p-value < 0.05 and the gene expressed in at least 10% of cells in one of the conditions.

Seurat analysis of WT vs ZC3H20 KO *Trypanosoma brucei* dataset

The ‘SS A’ and ‘SS B’ clusters were removed from the integrated dataset and the FindMarkers function was performed for all the four slender clusters between the WT and ZC3H20 KO conditions. A gene identified as being DE if the absolute Log2FC was more than 0.5, bonferroni corrected p-value < 0.05 and the gene expressed in at least 10% of cells in one of the conditions.

TradeSeq analysis of WT vs ZC3H20 KO *Trypanosoma 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 9. 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 Log2FC was greater than 0.5, bonferroni 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.

For all cases, genes were analysed in TriTrypDB to get their gene short names and functions (50).

WT vs Bcl11b KO *in vitro* T cell development dataset

The RDS (R dataset) file containing the scRNA-seq information was acquired from the authors (10).

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

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 V4 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

702 was used to determine the number of principal components (PC) used to create a UMAP
703 space embedding. Clustering was performed using the same number of PCs for UMAP
704 embedding calculation. Clustree (51) was used to help choose a resolution for clustering that
705 leads to stable clusters. Clusters were annotated as follows. Clusters that did not express
706 Cd3 genes but expressed Cd34 were defined as 'Cd34_TSP', clusters which did not express
707 Cd34 or Cd3 were characterized as 'early_T', clusters which had low expression of Cd3 but
708 no Cd34 expression were characterised as 'Cd3_lo_T', those with high Cd3 expression were
709 'Cd3_T'. Clusters which had high expression of Cd3 and expressed Ptcra and Rag1 were
710 defined as 'rearrange_TCR' and those that expressed the TCR signal transduction molecule
711 Zap70 and some cells that express the gene Trac, were defined as 'signaling_T'. Clusters
712 that express Rora were defined as 'Rora_T', clusters that expressed high levels of interferon
713 associated genes were defined as 'Interferon_response' and cells which clustered out with
714 the main body of the UMAP were classed as 'Outlier'. These three clusters were removed
715 before downstream analysis with TI, TrAGEDy, Seurat DE or TradeSeq.
716

717 TrAGEDy analysis of WT vs Bcl11b KO *in vitro* T cell development 718 dataset.

719 In order to build a gene feature space to construct PHATE embeddings and assess
720 dissimilarity with TrAGEDy, the top 250 DE genes (in terms of Log2FC) for each annotated
721 cluster for each of the four datasets were extracted. To reduce the effect of the cell cycle,
722 genes supplied by Seurat as being important in the cell cycle and genes included in the MGI
723 GO term 'cell cycle' were removed from the top 250 DE genes extracted, leaving 952 genes.
724 For all four of the individual datasets, 10 dimensional phate embeddings were generated
725 from the normalised gene expression count matrix for the features selected previously. For
726 the WT1 dataset, cells with a pseudotime higher than 0.1 (11 cells) were removed. TI was
727 then carried out using Slingshot on the PHATE embeddings with the Cd34_TSP cluster
728 chosen as the starting point. TrAGEDy was carried out on each of the conditions, using 100

729 interpolated points with a window size of the maximum pseudotime value, from either of the
730 two datasets being analysed, divided by 90. TrAGEDy aligned pseudotime values were then
731 created, resulting in a TrAGEDy aligned WT and a TrAGEDy aligned Bcl11b KO dataset.
732
733 The WT and Bcl11b KO TrAGEDy aligned datasets were then analysed using TrAGEDy.
734 TrAGEDy was carried out using 100 interpolated points with a window size of the maximum
735 pseudotime value, from either of the two datasets being analysed, divided by 90. TrAGEDy
736 differential expression test was then carried out across 7 windows of comparison. A gene
737 was returned as being differentially expressed if the Log2FC < 0.5, Bonferroni adjusted p-
738 value < 0.05 and the gene was expressed in at least 10% of cells in the window of
739 comparison, in one of the conditions. The final TrAGEDy aligned WT and Bcl11b KO
740 datasets were then analysed independently by TradeSeq to create the smooth gene
741 expression plots across the aligned pseudotime.
742

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

744 Using Seurat V4 integration (52–54), the WT condition dataset was integrated based on their
745 sequencing run with the same carried out for the Bcl11b KO. The top 2000 integration
746 anchor genes were identified by Seurat and used as the basis of the integrated expression
747 space. When the data was scaled, the effects of the cell cycle were regressed out by
748 passing the cell cycle associated genes Seurat provides into the var.to.regress parameter.
749 The resulting two integrated datasets were then integrated together to create a final
750 integrated Bcl11b KO/WT dataset. The top 3000 integration anchor genes were identified by
751 Seurat and used as the basis of the integrated expression space.
752
753 The integrated gene expression values were then scaled, and the cell cycle regressed out as
754 done previously. PCA was carried out on the scaled-regressed expression matrix, reducing
755 the dimensions down to 50. An elbow plot was used to determine the number of principal

756 components (PC) used to create a UMAP space embedding. Clustering was performed
757 using the same number of PCs for UMAP embedding calculation. Clustree was used to help
758 choose a resolution for clustering that leads to stable clusters and Clusters were annotated
759 as described previously. The FindMarkers function was used to identify DE genes between
760 the conditions for each cluster. Genes were determined to be DE if the absolute Log2FC
761 was greater than 0.5, bonferroni p-value < 0.05 and the gene was expressed in at least 10%
762 of the cells in one of the conditions for the cluster being compared.
763

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

766 Using the Seurat integrated, cell cycle regressed dataset; PHATE was carried out, using the
767 scaled and cell cycle regressed gene expression matrix. Slingshot was then applied to the
768 dataset, using the Cd34_TSP cluster as the starting point of the trajectory. The optimal
769 number of knots for fitting the general additive model to the dataset was assessed through
770 the evaluateK function, with the optimal number determined to be 8. fitGAM was run on the
771 dataset followed by the conditionTest function to identify DE genes between the WT and
772 Bcl11b KO conditions. 7 different knot comparisons were made, and a gene were
773 determined to be DE if the Log2FC was greater than 0.5, bonferroni p-value < 0.05 and the
774 gene was expressed in at least 10% of the cells that fell within the current knot comparison,
775 in one of the conditions.
776

777 Runtime experiments

778 Packages were run single threaded on a virtual machine with 70 gigabytes of allotted RAM
779 and 3 GHz clock speed. Runtime experiment code detailing what sections were included for
780 each package when calculating runtime can be found on our GitHub.
781

782 Package versions details

All analysis and experiments were carried out on R 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 4.2.0, Dyngen version 1.0.5, PhateR version 1.0.7, Slingshot version 2.2.1, Single Cell Experiment version 1.16.0, TradeSeq version 1.6.0 were used.

References

1. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol [Internet]. 2014;32(4):381–6. Available from: https://doi.org/10.1038/nbt.2859
2. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nat Biotechnol [Internet]. 2019;37(5):547–54. Available from: https://doi.org/10.1038/s41587-019-0071-9
3. van den Berge K, Roux de Bézieux H, Street K, Saelens W, Cannoodt R, Saeys Y, et al. Trajectory-based differential expression analysis for single-cell sequencing data. Nat Commun [Internet]. 2020;11(1):1201. Available from: https://doi.org/10.1038/s41467-020-14766-3
4. Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. Nat Methods [Internet]. 2017;14(3):309–15. Available from: https://doi.org/10.1038/nmeth.4150
5. Song D, Li JJ. PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell RNA sequencing data. Genome Biol [Internet]. 2021;22(1):124. Available from: https://doi.org/10.1186/s13059-021-02341-y
6. Cano-Gamez E, Soskic B, Roumeliotis TI, So E, Smyth DJ, Baldrighi M, et al. Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4+ T cells to cytokines. Nat Commun [Internet]. 2020;11(1):1801. Available from: https://doi.org/10.1038/s41467-020-15543-y
7. Briggs EM, Rojas F, McCulloch R, Matthews KR, Otto TD. Single-cell transcriptomic analysis of bloodstream Trypanosoma brucei reconstructs cell cycle progression and developmental quorum sensing. Nat Commun [Internet]. 2021;12(1):5268. Available from: https://doi.org/10.1038/s41467-021-25607-2
8. Alpert A, Moore LS, Dubovik T, Shen-Orr SS. Alignment of single-cell trajectories to compare cellular expression dynamics. Nat Methods [Internet]. 2018;15(4):267–70. Available from: https://doi.org/10.1038/nmeth.4628
9. Cannoodt R, Saelens W, Deconinck L, Saeys Y. Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells. Nat Commun [Internet]. 2021;12(1):3942. Available from: https://doi.org/10.1038/s41467-021-24152-2
10. Zhou W, Gao F, Romero-Wolf M, Jo S, Rothenberg E v. Single-cell deletion analyses show control of pro–T cell developmental speed and pathways by Tcf7, Spi1, Gata3, Bcl11a, Erg, and Bcl11b. Sci Immunol [Internet]. 2022 Dec 21;7(71):eabm1920. Available from: https://doi.org/10.1126/sciimmunol.abm1920

11. Seo W, Taniuchi I. Transcriptional regulation of early T-cell development in the thymus. Eur J Immunol [Internet]. 2016 Mar 1;46(3):531–8. Available from: https://doi.org/10.1002/eji.201545821
12. Hwang JR, Byeon Y, Kim D, Park SG. Recent insights of T cell receptor-mediated signaling pathways for T cell activation and development. Exp Mol Med [Internet]. 2020;52(5):750–61. Available from: https://doi.org/10.1038/s12276-020-0435-8
13. Mombaerts P, Iacomini J, Johnson RS, Herrup K, Tonegawa S, Papaioannou VE. RAG-1-deficient mice have no mature B and T lymphocytes. Cell [Internet]. 1992 Mar 6;68(5):869–77. Available from: https://doi.org/10.1016/0092-8674(92)90030-G
14. Moon KR, van Dijk D, Wang Z, Gigante S, Burkhardt DB, Chen WS, et al. Visualizing structure and transitions in high-dimensional biological data. Nat Biotechnol [Internet]. 2019;37(12):1482–92. Available from: https://doi.org/10.1038/s41587-019-0336-3
15. Giotti B, Chen SH, Barnett MW, Regan T, Ly T, Wiemann S, et al. Assembly of a parts list of the human mitotic cell cycle machinery. J Mol Cell Biol [Internet]. 2019 Aug 19;11(8):703–18. Available from: https://doi.org/10.1093/jmcb/mjy063
16. Yuniati L, Scheijen B, van der Meer LT, van Leeuwen FN. Tumor suppressors BTG1 and BTG2: Beyond growth control. J Cell Physiol [Internet]. 2019 May 1;234(5):5379–89. Available from: https://doi.org/10.1002/jcp.27407
17. Rashida Gnanaprakasam JN, Wu R, Wang R. Metabolic Reprogramming in Modulating T Cell Reactive Oxygen Species Generation and Antioxidant Capacity. Front Immunol [Internet]. 2018;9. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2018.01075
18. Liau NPD, Laktyushin A, Lucet IS, Murphy JM, Yao S, Whitlock E, et al. The molecular basis of JAK/STAT inhibition by SOCS1. Nat Commun [Internet]. 2018;9(1):1558. Available from: https://doi.org/10.1038/s41467-018-04013-1
19. Grujic M, Braga T, Lukinius A, Eloranta ML, Knight SD, Pejler G, et al. Serglycin-deficient Cytotoxic T Lymphocytes Display Defective Secretory Granule Maturation and Granzyme B Storage*. Journal of Biological Chemistry [Internet]. 2005;280(39):33411–8. Available from: https://www.sciencedirect.com/science/article/pii/S0021925820790708
20. Korinfskaya S, Parameswaran S, Weirauch MT, Barski A. Runx Transcription Factors in T Cells—What Is Beyond Thymic Development? Front Immunol [Internet]. 2021;12. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2021.701924
21. Schilham MW, Moerer P, Cumano A, Clevers HC. Sox-4 facilitates thymocyte differentiation. Eur J Immunol [Internet]. 1997 May 1;27(5):1292–5. Available from: https://doi.org/10.1002/eji.1830270534
22. B RD. V(D)J Recombination: Mechanism, Errors, and Fidelity. Microbiol Spectr [Internet]. 2014 Nov 21;2(6):2.6.18. Available from: https://doi.org/10.1128/microbiolspec.MDNA3-0041-2014
23. Moran M, Miceli MC. Engagement of GPI-Linked CD48 Contributes to TCR Signals and Cytoskeletal Reorganization: A Role for Lipid Rafts in T Cell Activation. Immunity [Internet]. 1998;9(6):787–96. Available from: https://www.sciencedirect.com/science/article/pii/S1074761300806445
24. Lobach DF, Hensley LL, Ho W, Haynes BF. Human T cell antigen expression during the early stages of fetal thymic maturation. J Immunol. 1985 Sep;135(3):1752–9.
25. Aandahl EM, Sandberg JK, Beckerman KP, Taskén K, Moretto WJ, Nixon DF. CD7 Is a Differentiation Marker That Identifies Multiple CD8 T Cell Effector Subsets1. The

871 Journal of Immunology [Internet]. 2003 Mar 1;170(5):2349–55. Available from:
872 https://doi.org/10.4049/jimmunol.170.5.2349
873 26. Agresta L, Hoebe KHN, Janssen EM. The Emerging Role of CD244 Signaling in
874 Immune Cells of the Tumor Microenvironment. Front Immunol [Internet]. 2018;9.
875 Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2018.02809
876 27. Dong K, Chen W, Pan X, Wang H, Sun Y, Qian C, et al. FCER1G positively relates to
877 macrophage infiltration in clear cell renal cell carcinoma and contributes to
878 unfavorable prognosis by regulating tumor immunity. BMC Cancer [Internet].
879 2022;22(1):140. Available from: https://doi.org/10.1186/s12885-022-09251-7
880 28. Reuner B, Vassella E, Yutzy B, Boshart M. Cell density triggers slender to stumpy
881 differentiation of Trypanosoma brucei bloodstream forms in culture. Mol Biochem
882 Parasitol [Internet]. 1997;90(1):269–80. Available from:
883 https://www.sciencedirect.com/science/article/pii/S0166685197001606
884 29. Rojas F, Silvester E, Young J, Milne R, Tettey M, Houston DR, et al. Oligopeptide
885 Signaling through TbGPR89 Drives Trypanosome Quorum Sensing. Cell [Internet].
886 2019;176(1):306-317.e16. Available from:
887 https://www.sciencedirect.com/science/article/pii/S0092867418313977
888 30. Matthews KR. Trypanosome Signaling—Quorum Sensing. Annu Rev Microbiol
889 [Internet]. 2021 Oct 8;75(1):495–514. Available from: https://doi.org/10.1146/annurev-
890 micro-020321-115246
891 31. Liu B, Kamanyi Marucha K, Clayton C. The zinc finger proteins ZC3H20 and ZC3H21
892 stabilise mRNAs encoding membrane proteins and mitochondrial proteins in insect-
893 form Trypanosoma brucei. Mol Microbiol [Internet]. 2020 Feb 1;113(2):430–51.
894 Available from: https://doi.org/10.1111/mmi.14429
895 32. Cayla M, McDonald L, MacGregor P, Matthews K. An atypical DYRK kinase connects
896 quorum-sensing with posttranscriptional gene regulation in Trypanosoma brucei.
897 Engstler M, Soldati-Favre D, Acosta Serrano Á, editors. Elife [Internet].
898 2020;9:e51620. Available from: https://doi.org/10.7554/eLife.51620
899 33. Mantilla BS, Marchese L, Casas-Sánchez A, Dyer NA, Ejeh N, Biran M, et al. Proline
900 Metabolism is Essential for Trypanosoma brucei brucei Survival in the Tsetse Vector.
901 PLoS Pathog [Internet]. 2017 Jan 23;13(1):e1006158-. Available from:
902 https://doi.org/10.1371/journal.ppat.1006158
903 34. Marcus JP, Dekker EE. Threonine formation via the coupled activity of 2-amino-3-
904 ketobutyrate coenzyme A lyase and threonine dehydrogenase. J Bacteriol. 1993
905 Oct;175(20):6505–11.
906 35. Fletcher DA, Mullins RD. Cell mechanics and the cytoskeleton. Nature [Internet].
907 2010;463(7280):485–92. Available from: https://doi.org/10.1038/nature08908
908 36. Gupta SK, Chikne V, Eliaz D, Tkacz ID, Naboishchikov I, Carmi S, et al. Two splicing
909 factors carrying serine-arginine motifs, TSR1 and TSR1IP, regulate splicing, mRNA
910 stability, and rRNA processing in Trypanosoma brucei. RNA Biol [Internet]. 2014 Jun
911 1;11(6):715–31. Available from: https://doi.org/10.4161/rna.29143
912 37. Tyler Weisbarth R, Das A, Castellano P, Fisher MA, Wu H, Bellofatto V. The
913 Trypanosoma cruzi RNA-binding protein RBP42 is expressed in the cytoplasm
914 throughout the life cycle of the parasite. Parasitol Res [Internet]. 2018;117(4):1095–
915 104. Available from: https://doi.org/10.1007/s00436-018-5787-9
916 38. Claudia H, Corinna B, Stefanie B, Louise E, Van-Duc L, Mhairi S, et al. Small
917 Trypanosome RNA-Binding Proteins TbUBP1 and TbUBP2 Influence Expression of F-

Box Protein mRNAs in Bloodstream Trypanosomes. Eukaryot Cell [Internet]. 2007 Nov 1;6(11):1964–78. Available from: https://doi.org/10.1128/EC.00279-07

39. Dean S, Marchetti R, Kirk K, Matthews KR. A surface transporter family conveys the trypanosome differentiation signal. Nature [Internet]. 2009;459(7244):213–7. Available from: https://doi.org/10.1038/nature07997

40. Li Y, Li K, Zhu L, Li B, Zong D, Cai P, et al. Development of double-positive thymocytes at single-cell resolution. Genome Med [Internet]. 2021;13(1):49. Available from: https://doi.org/10.1186/s13073-021-00861-7

41. Allen CE, Mak C ho, Wu LC. The κB transcriptional enhancer motif and signal sequences of V(D)J recombination are targets for the zinc finger protein HIVEP3/KRC: a site selection amplification binding study. BMC Immunol [Internet]. 2002;3(1):10. Available from: https://doi.org/10.1186/1471-2172-3-10

42. Peng HY, Lucavs J, Ballard D, Das JK, Kumar A, Wang L, et al. Metabolic Reprogramming and Reactive Oxygen Species in T Cell Immunity. Front Immunol [Internet]. 2021;12. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2021.652687

43. Geigy R, Huber M, Weinman D, Wyatt GR. Demonstration of trehalose in the vector of African trypanosomiasis: the tsetse fly. Acta Trop. 1959;16:255–62.

44. Lamour N, Rivière L, Coustou V, Coombs GH, Barrett MP, Bringaud F. Proline Metabolism in Procyclic Trypanosoma brucei Is Down-regulated in the Presence of Glucose*. Journal of Biological Chemistry [Internet]. 2005;280(12):11902–10. Available from: https://www.sciencedirect.com/science/article/pii/S0021925820808990

45. Millerioux Y, Ebikeme C, Biran M, Morand P, Bouyssou G, Vincent IM, et al. The threonine degradation pathway of the Trypanosoma brucei procyclic form: the main carbon source for lipid biosynthesis is under metabolic control. Mol Microbiol [Internet]. 2013 Oct 1;90(1):114–29. Available from: https://doi.org/10.1111/mmi.12351

46. Das A, Morales R, Banday M, Garcia S, Hao L, Cross GAM, et al. The essential polysome-associated RNA-binding protein RBP42 targets mRNAs involved in *Trypanosoma brucei* energy metabolism. RNA. 2012 Nov;18(11):1968–83.

47. Wagner DE, Klein AM. Lineage tracing meets single-cell omics: opportunities and challenges. Nat Rev Genet [Internet]. 2020;21(7):410–27. Available from: https://doi.org/10.1038/s41576-020-0223-2

48. Schraivogel D, Gschwind AR, Milbank JH, Leonce DR, Jakob P, Mathur L, et al. Targeted Perturb-seq enables genome-scale genetic screens in single cells. Nat Methods [Internet]. 2020;17(6):629–35. Available from: https://doi.org/10.1038/s41592-020-0837-5

49. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC Genomics [Internet]. 2018;19(1):477. Available from: https://doi.org/10.1186/s12864-018-4772-0

50. Amos B, Aurrecoechea C, Barba M, Barreto A, Basenko EY, Bażant W, et al. VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. Nucleic Acids Res [Internet]. 2022 Jan 7;50(D1):D898–911. Available from: https://doi.org/10.1093/nar/gkab929

51. Zappia L, Oshlack A. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. Gigascience [Internet]. 2018 Jul 1;7(7):giy083. Available from: https://doi.org/10.1093/gigascience/giy083

52. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol [Internet]. 2018;36(5):411–20. Available from: https://doi.org/10.1038/nbt.4096
53. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell [Internet]. 2021;184(13):3573-3587.e29. Available from: https://www.sciencedirect.com/science/article/pii/S0092867421005833
54. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive Integration of Single-Cell Data. Cell [Internet]. 2019;177(7):1888-1902.e21. Available from: https://www.sciencedirect.com/science/article/pii/S0092867419305598

Acknowledgements

This research was funded by the Medical Research Council through their Precision Medicine DTP (MRC grant number: MR/N013166/1). This research was funded in whole, or in part, by the Wellcome Trust [Grant numbers 218648/Z/19/Z to EMB, 104111/Z/14/ZR to TDO/RM, 206815/Z/17/Z & 224501/Z/21/Z to RM, and 103740/Z14/Z to KRM] and the BBSRC Project grants (BB/R017166/1 & BB/W001101/1 to RM).

Availability of data and materials

The functions to run TrAGEDy are available on our GitHub (https://github.com/No2Ross/TrAGEDy) along with the code required to repeat our analysis.

Author Information

Authors and Affiliations

Wellcome Centre for Integrative Parasitology, Institute of Infection, Immunity and Inflammation, University of Glasgow, Glasgow, UK

Ross F. Laidlaw, Emma M. Briggs, Richard McCulloch & Thomas D. Otto

Contributions

R.F.L. and T.D.O. conceived the project and R.F.L. implemtented the TrAGEDy package. R.F.L., T.D.O., E.M.B. and R.M. designed experiments. R.F.L. performed the experiments and collected data. R.F.L., T.D.O., E.M.B. and R.M. analyzed the data. R.F.L. wrote the manuscript with editing from R.M., E.M.B., K.R.M. and T.D.O. All authors read, edited and approved the manuscript.

Corresponding authors

Correspondence to Ross F. Laidlaw or Thomas D. Otto.

Ethics declarations

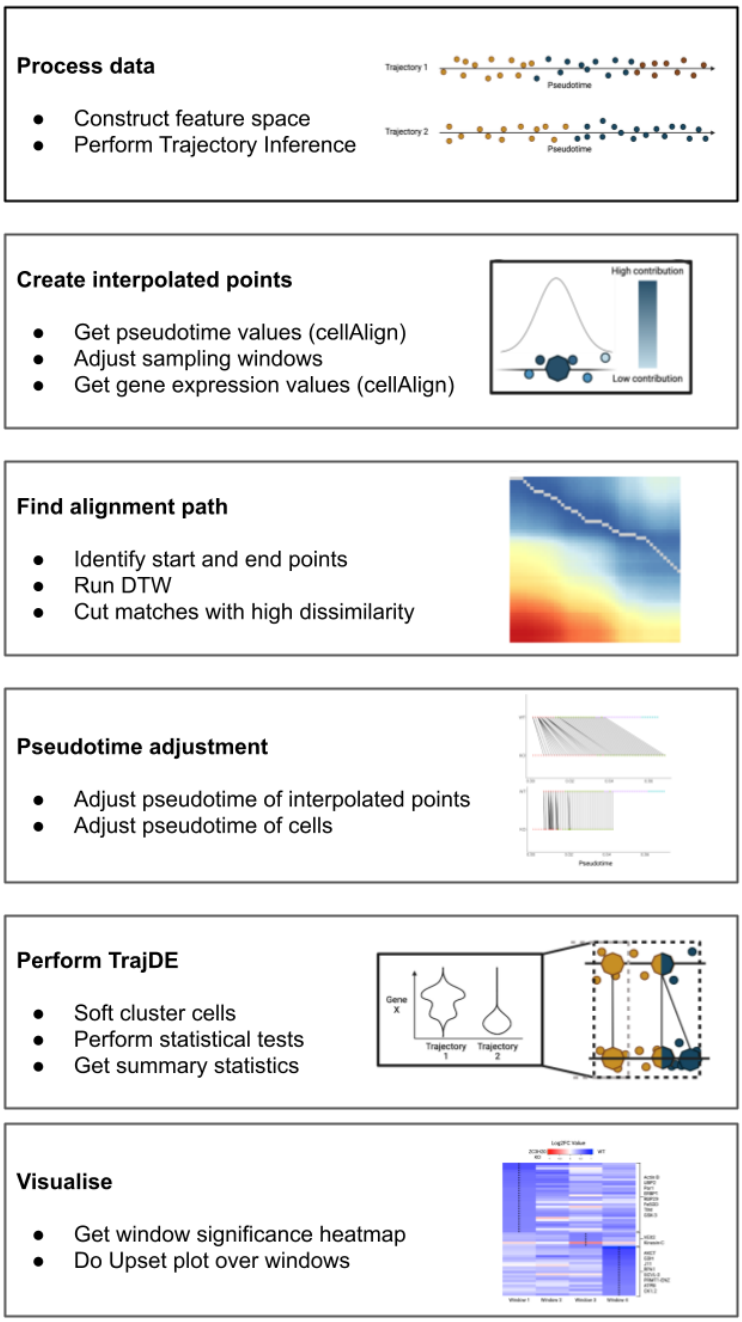
Ethics approval and consent to participate

Not applicable.

Competing interests

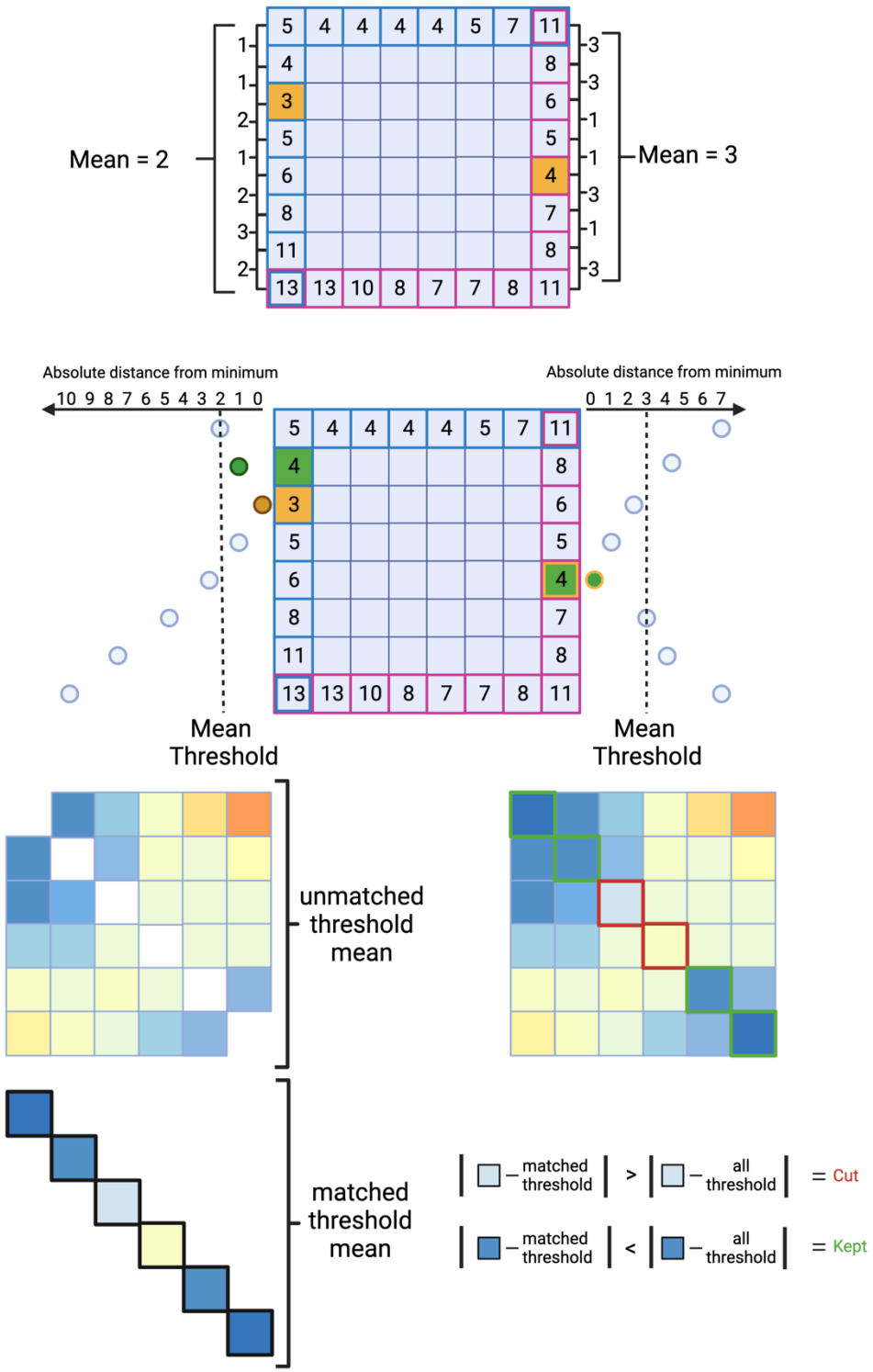
The authors declared that they have no competing interests.

Supplementary figures and tables



Supplementary figure 1. Workflow of TrAGEDy process.

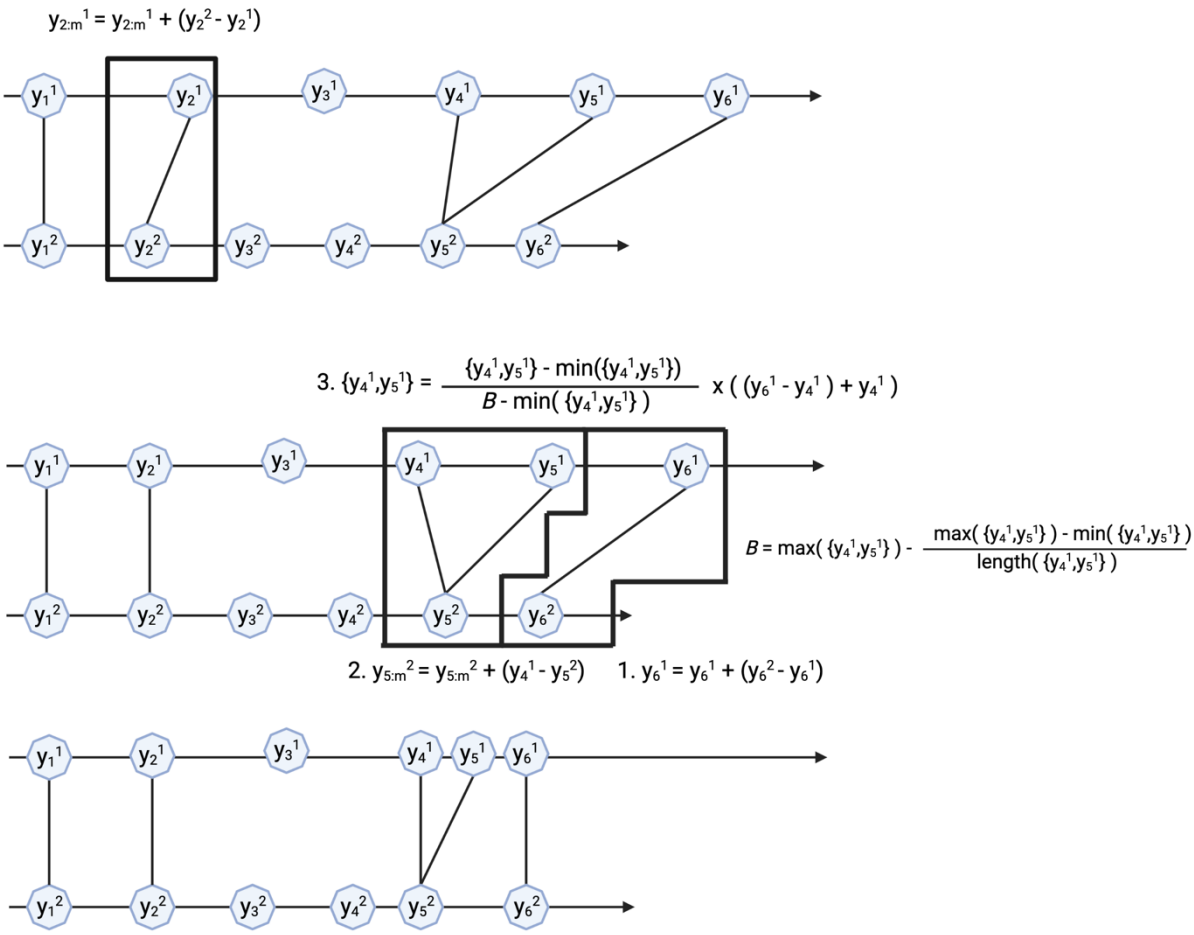
Workflow detailing the different parts of the TrAGEDy process. Sections which were derived from cellAlign are noted in the process boxes.



Supplementary figure 2. Graphical explanation of the TrAGEDy process of modifying the penalty matrix.

TrAGEDy finds the minimum dissimilarity score index from the beginning and end of the matrix from either the row or column. The absolute change in dissimilarity score across the

row/column is captured and the median or mean threshold of change is found (A). To identify whether a new index will become the start/end point we find the absolute difference between each index and the current start/end point and identify a new start/end point if it lies below the threshold (B). In order to cut matches from the DTW aligned path, a mean/median threshold for the unmatched and matched indexes are calculated and if an index in the DTW path is closer to the unmatched threshold than the matched, it is cut from the path (C).

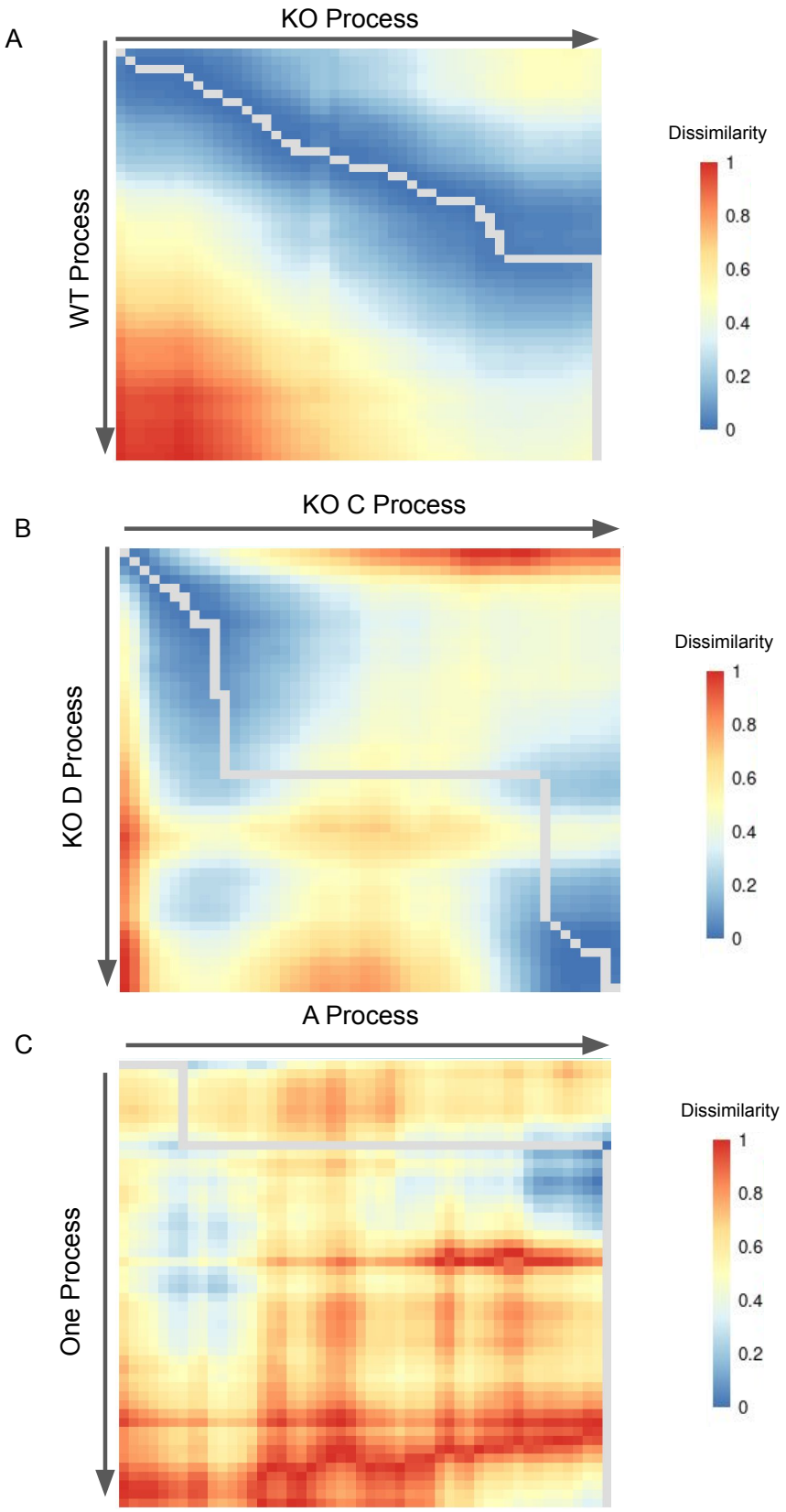


Supplementary figure 3. Graphical explanation of the alignment of interpolated point pseudotime values, from the DTW calculated path.

There are two types of movement that can occur when aligning pseudotime values, an individual match (A) or a multi-match (B). 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 (C).

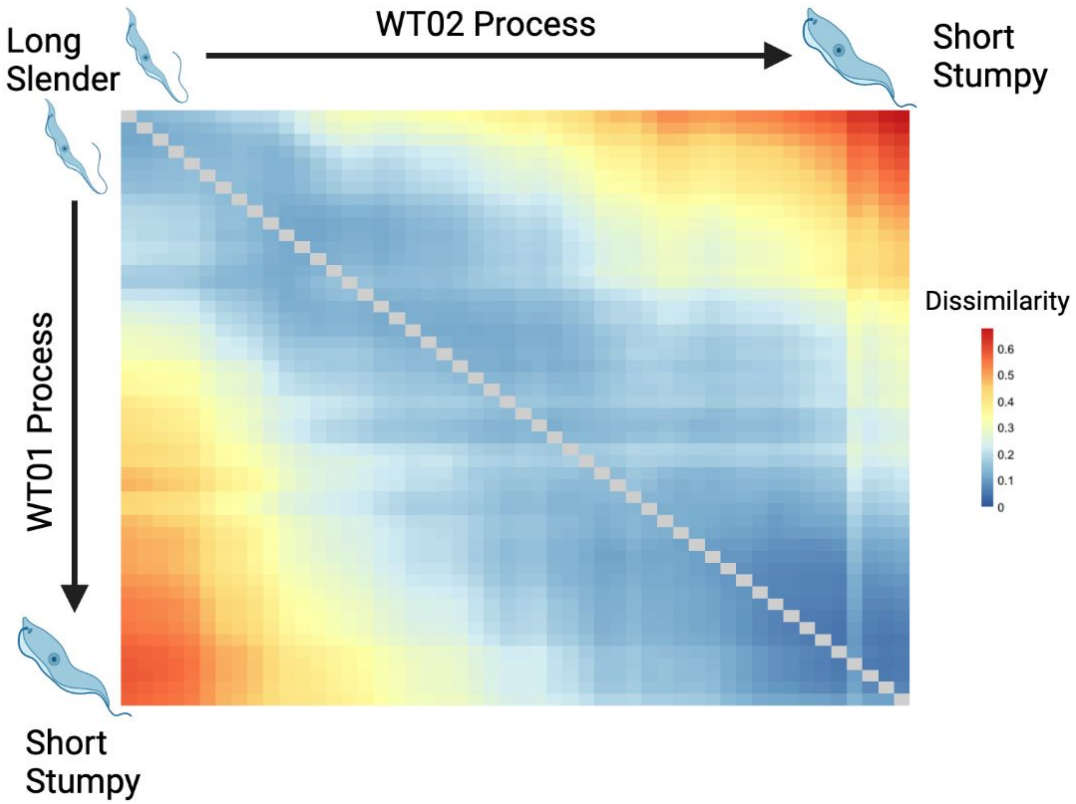
1039


A
KO Process
WT Process
Dissimilarity
1
0.8
0.6
0.4
0.2
0
B
KO C Process
KO D Process
Dissimilarity
1
0.8
0.6
0.4
0.2
0
C
A Process
One Process
Dissimilarity
1
0.8
0.6
0.4
0.2
0

1040

Supplementary figure 4. cellAlign alignments of simulated datasets using Pearson correlation dissimilarity score.

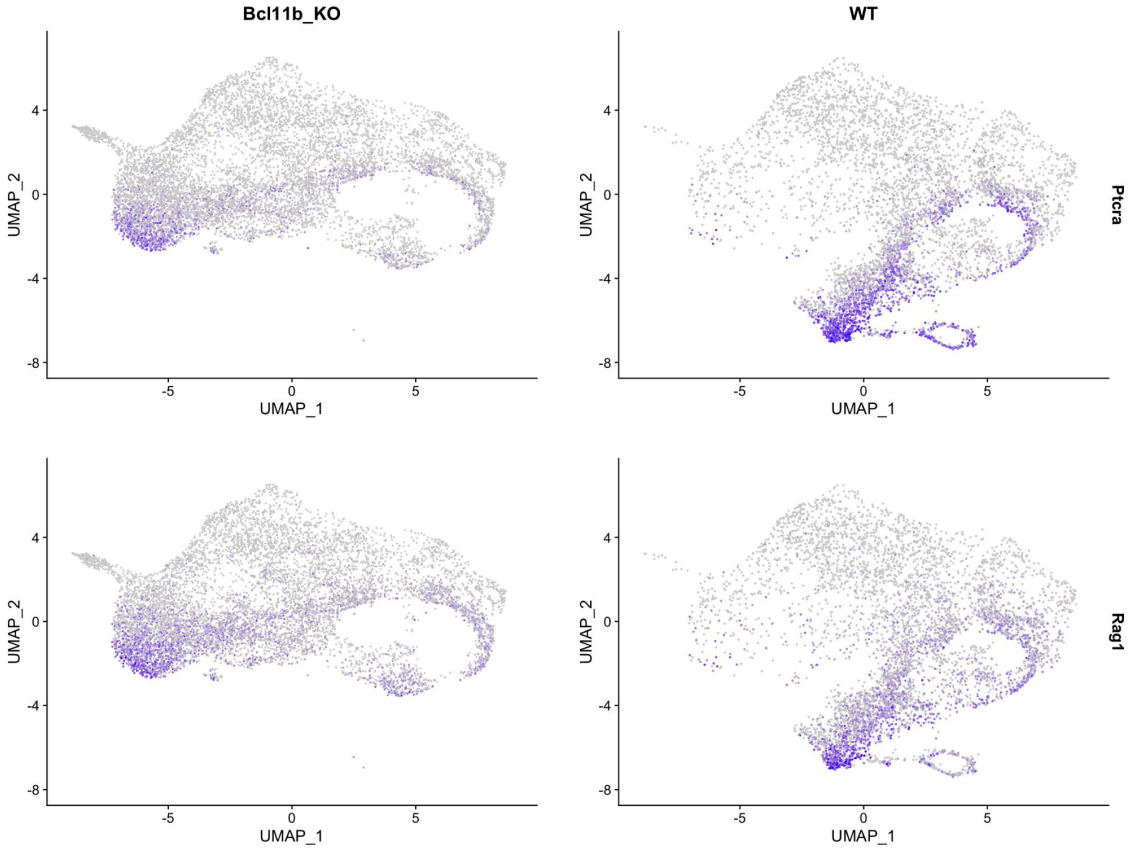
The alignments of the three simulated datasets experiments as calculated by cellAlign. WT vs KO started shared simulation (A), KO C vs KO D diverge to converge simulation (B), negative control A vs One simulation (C). Pearson correlation was used to assess transcriptional dissimilarity between the conditions.



Supplementary figure 5. TrAGEDy alignment of biological replicates of WT slender to stumpy *Trypanosoma brucei* transition.

TrAGEDy alignment of the WT01 and WT02 slender to stumpy *Trypanosoma 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.

1057



1058

1059 **Supplementary figure 6. Expression of markers associated with DN3 stage of T cell**

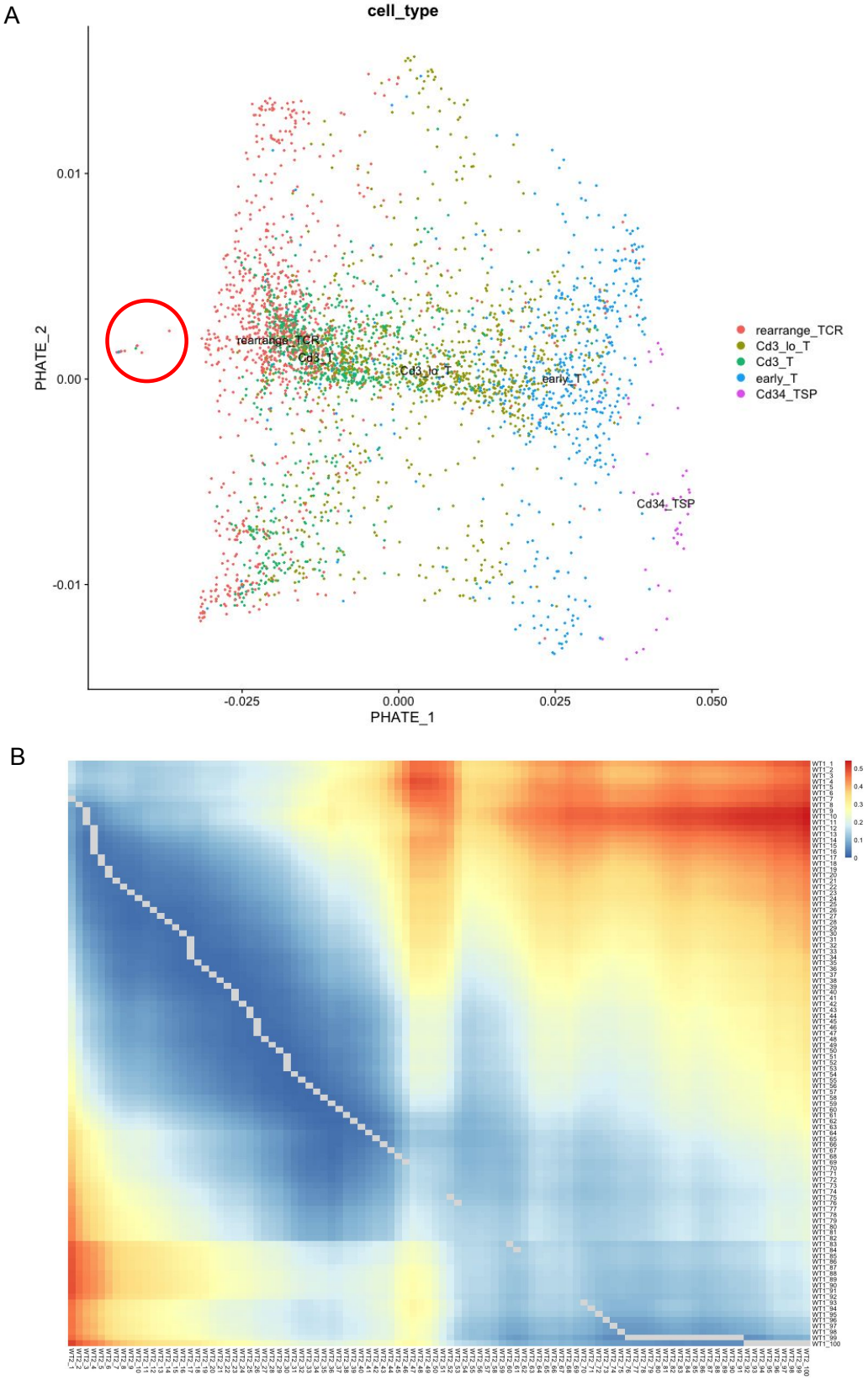
1060 **development in Bcl11b KO and WT T cells.**

1061 UMAP showing the normalised gene expression levels of markers of double negative 3

1062 (DN3) stage of T cell development: Ptcra (pre-TCR alpha chain) and Rag1 (recombination

1063 activating gene 1) in T cells under WT (left) and Bcl11b KO (right) conditions.

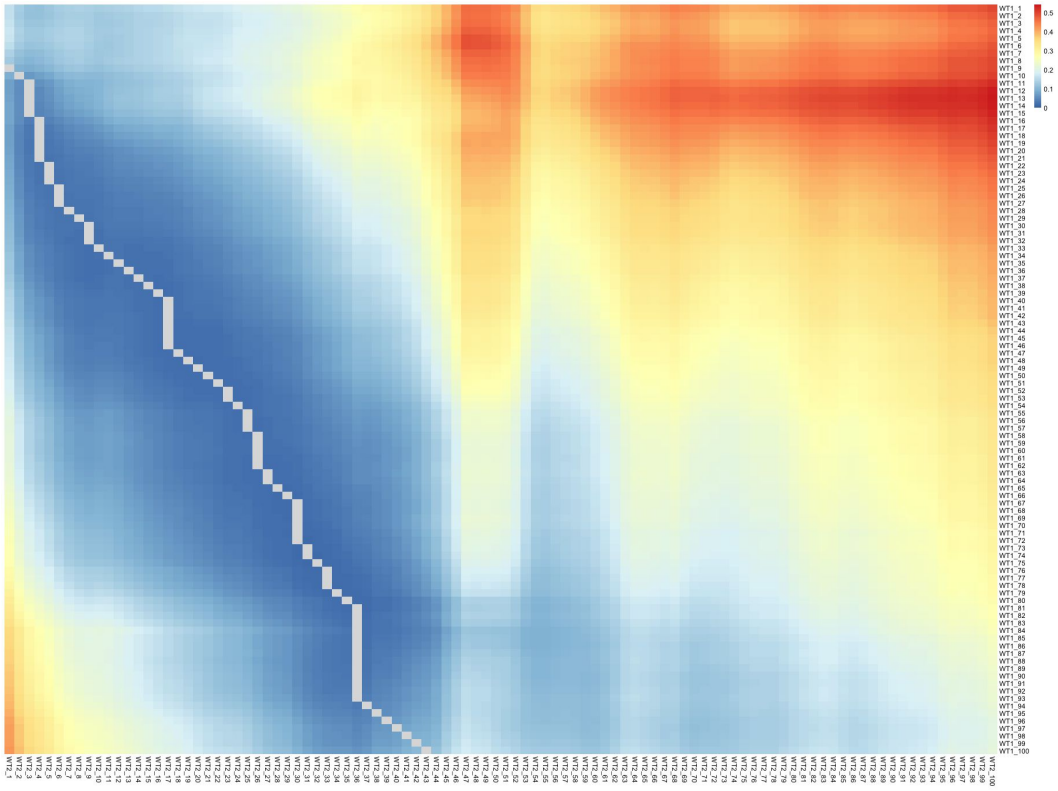
1064



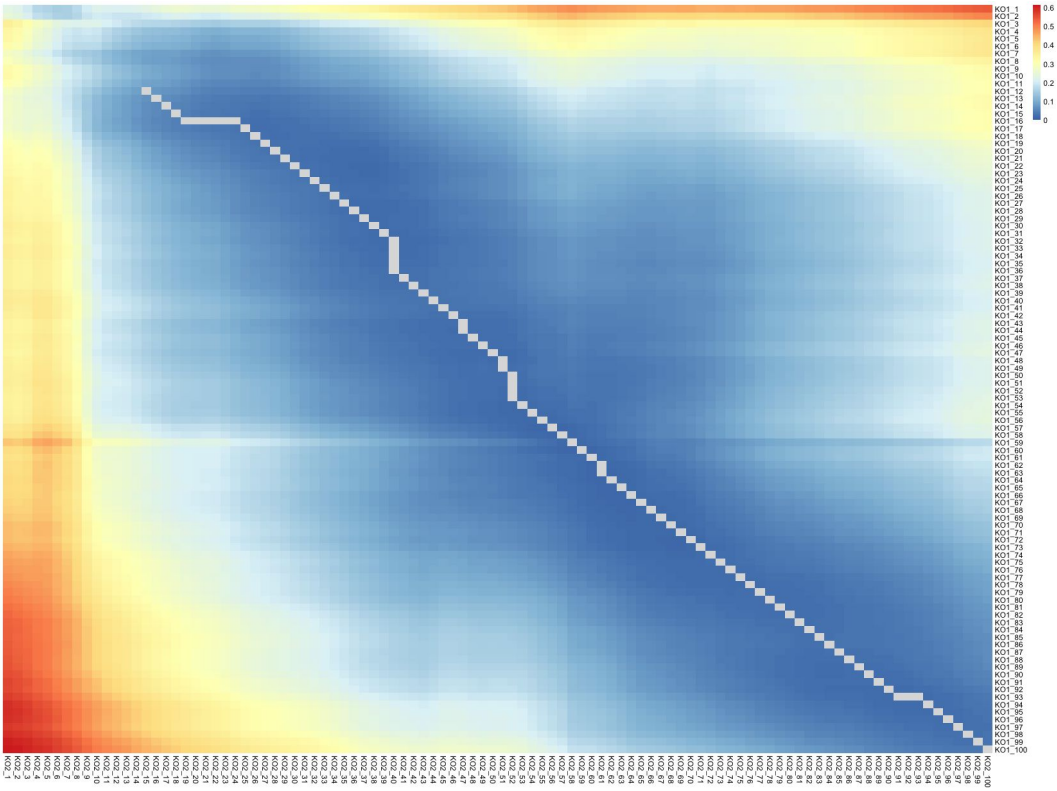
Supplementary Figure 7. Impact of outlier cells on TrAGEDy alignments.

1068 Plot showing the PHATE embeddings of the WT1 T cell dataset with each cell coloured by
1069 its cell type annotation. Circle identifies cells which are separate from the main body of the
1070 trajectory (A). TrAGEDy alignment of the WT1 and WT2 T cell datasets when the cells
1071 circled in A are kept in the trajectory. Dissimilarity in gene expression of interpolated points
1072 was calculated using Spearman correlation with blue meaning low dissimilarity and red
1073 meaning high dissimilarity.

A



B



Supplementary figure 8. 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.

Supplementary table information:

Sup table 1 – the DE genes and associated statistics for TrAGEDy applied to T cell dataset for 0.25 and 0.5 Log2FC

Sup table 2 - the DE genes and associated statistics for TradeSeq applied to T cell dataset for 0.25 and 0.5 Log2FC

Sup table 3 - the DE genes and associated statistics for Seurat applied to T cell dataset for 0.25 and 0.5 Log2FC

Sup table 4 – the DE genes and associated statistics for TrAGEDy applied to *T. brucei* dataset for 0.25 and 0.5 Log2FC

Sup table 5 - the DE genes and associated statistics for TradeSeq applied to *T. brucei* dataset for 0.25 and 0.5 Log2FC

Sup table 6 - the DE genes and associated statistics for Seurat applied to *T. brucei* dataset for 0.25 and 0.5 Log2FC

Sup table 7 – full list of genes in heatmap for T cell and *T. brucei* analysis