A single-parasite transcriptional landscape of asexual development in Toxoplasma gondii

Yuan Xue[1], Terence C. Theisen[2], Suchita Rastogi[2], Abel Ferrel[2], Stephen R. Quake[1,3,4,*], John C. Boothroyd[2,*]

[1]Department of Bioengineering, [3]Department of Applied Physics, Stanford University, Stanford, CA, USA

[2]Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA, USA

[4]Chan Zuckerberg Biohub, San Francisco, CA, USA

*Correspondence should be addressed to J.C.B. (jboothr@stanford.edu) and S.R.Q. (steve@quake-lab.org)

Abstract (150 words)

Toxoplasma gondii, a protozoan parasite, undergoes a complex and poorly understood developmental process that is critical for completing its intricate life cycle, including establishing a chronic infection in its intermediate hosts. Here, we applied single-cell RNA-sequencing (scRNA-seq) to ≥ 5,000 *Toxoplasma* at single-parasite resolution in tachyzoite and bradyzoite stages using three widely studied strains. We resolve the oscillatory nature of cell cycle progression in an asynchronized population of the type I strain, RH. Using scRNA-seq, we also construct a comprehensive atlas of asexual development and cell-cycle in the Type II strains, Pru and ME49, revealing hidden heterogeneity in the course of development and transcription factors associated with each developmental state. Lastly, we combined projection scoring with noise analysis to show that the expression of a subset of parasite-specific genes, including ones that encode surface antigens, varies independently of measurement noise, cell cycle, and asexual development. Overall, our results reveal an unprecedented and surprising level of heterogeneity in *Toxoplasma gondii* and provide a molecular resource for understanding protozoan parasite development.

Keywords: single-cell RNA sequencing, protozoa, cell cycle, Toxoplasma, development

Main text (5000 words)

Introduction

Toxoplasma gondii is an intracellular protozoan parasite that is thought to infect over a quarter of the world's population[1]. Like some of its *Apicomplexan* cousins, *Toxoplasma* undergoes a complex developmental transition inside the host. In intermediate hosts, including humans and virtually all other non-feline warm-blooded animals, *Toxoplasma* parasites remain haploid and transition from a replicative, virulent tachyzoite to an encysted, quasi-dormant bradyzoite. This asexual developmental transition is tightly coupled to the clinical progression of *Toxoplasma* infection. Although acute infection with tachyzoites produces few if any symptoms in healthy human children and adults, infected individuals, if left untreated, progress to a chronic stage wherein tachyzoites transition to bradyzoites that can persist for life in neurons and muscle cells. When infected individuals become immunocompromised, such as in chemotherapy, HIV infection, or organ transplantation[2,3] bradyzoites can reactivate to become tachyzoites, causing severe neurological damage and even death. While no causal link has been established, a population-wide study has uncovered significant association of *Toxoplasma* infection with risk for schizophrenia in chronically infected humans[4]. Chronic infection in mice has been observed to induce behavioral changes such as loss of aversion to cat urine, which is hypothesized to increase the transmission rate of *Toxoplasma* to its definitive feline host where sexual reproduction occurs[5]. As there are no therapeutic interventions to prevent or clear cysts in infected individuals, understanding how *Toxoplasma* transitions through its life stages remains of critical importance.

The development of *in vitro* methods to induce *Toxoplasma* differentiation have facilitated investigation of several aspects of chronic infection, including transition of tachyzoites to bradyzoites[6,7]. Bulk transcriptomic analyses of *Toxoplasma gondii* at distinct asexual stages reveal genetic modules that are expressed in each stage[8–15], including AP2 transcription factors that are thought to play a role in differentiation[16,17]; however, transitioning parasites convert to the bradyzoite stage asynchronously and display a high degree of heterogeneity along the developmental pathway and in gene expression[18,19]. Furthermore, parasites within the same tissue cysts have been shown to display heterogeneity in the expression of bradyzoite marker proteins[20]. The transition of tachyzoites to the bradyzoite stage results in an overwhelming majority of mature bradyzoites in the G_1 phase of the cell cycle that divide slowly, if at all[21,22]. Furthermore, tachyzoites exhibit slower growth kinetics immediately prior to the bradyzoite transition[21,23]. This suggests that parasites exit the cell cycle to differentiate into bradyzoites, a pattern consistent with developmental processes in several other eukaryotic organisms[24,25]. Dissecting these cell cycle aspects of stage conversion requires a more detailed analysis than has been possible with bulk measurement of

tachyzoite or bradyzoite populations, or with the use of genetically modified parasites coupled with chemical synchronization of cell cycle progression[13,26–28]. This is because the latter approaches require large quantities of synchronized parasites and can potentially introduce artificial perturbations. Furthermore, bulk measurement fails to distinguish parasite-to-parasite variation that is independent of cell cycle or known developmental processes, potentially missing the phenotypic diversity intrinsic to a population of cells.

Single-cell RNA sequencing (scRNA-seq) offers a powerful and unbiased approach to reveal the underlying heterogeneity in an asynchronous population of cells. Droplet and FACS-based approaches have already been applied towards multicellular parasites such as *Schistosoma* to reveal developmental changes within different hosts[29]. Recently, scRNA-seq has revealed a surprising degree of heterogeneity in another apicomplexan parasite, *Plasmodium*[30–32]. Analyses derived from these single-parasite measurements uncovered rare and critical transition events in parasite development that were undetectable in bulk measurements. Combined with novel analytical development and increases in measurement throughput, scRNA-seq is becoming a widely adopted tool for resolving cellular changes in a quantitative and system-wide fashion.

Here, we performed scRNA-seq to reconstruct transcriptional dynamics of asynchronous *Toxoplasma* parasites in the course of cell cycle and asexual development *in vitro*. We benchmarked the purity of isolation, as well as sensitivity and accuracy of our measurements, demonstrating that this experimental approach can isolate single parasites to resolve the transcriptional variation of biological processes. We show that cell cycle status can be accurately inferred from the transcriptional signatures of an asynchronous population of Type I (RH) tachyzoites at single-parasite resolution. Using Type II strains (Pru and ME49) switching to bradyzoites under alkaline induction, we resolved a comprehensive single-parasite atlas of asexual development together with cell cycle state annotation, identifying transcription factors that are associated with each developmental state and revealing previously hidden heterogeneity in the parasite. Furthermore, we identify a class of highly variable genes, including ones that encode surface antigens and dense granule effectors, which exhibit parasite-to-parasite variability that cannot be explained by either measurement noise, cell cycle, or asexual development. Our combined results suggest that this prevalent protozoan parasite may exhibit much greater heterogeneity than previously appreciated.

Results

Technical validation of single-parasite sorting and sequencing

There are more than a dozen approaches available for single-cell isolation and transcriptome amplification. Based on benchmark comparisons, Smart-seq2 generally has higher sensitivity than competing droplet-based approaches[33,34]. We reasoned that sensitive measurement is crucial in our study given that single *Toxoplasma gondii* parasites are at least 50-fold smaller in volume than a typical mammalian cell, and thus the average parasite gene is likely expressed with much lower copy number per cell than a typical mammalian gene. For our initial studies, we used the common Type I lab strain of *Toxoplasma*, RH, grown *in vitro* in human foreskin fibroblasts (HFFs). Following such growth, individual tachyzoites were released by passage through a narrow gauge needle and then purified by fluorescence activated cell sorting (FACS) into 384-well plates. We then synthesized, amplified, and barcoded cDNA using Smart-seq2 and Illumina Nextera protocols. Reaction in 384-well plates effectively reduced the reagent cost by four-fold compared to the 96-well format. The sequenced reads were bioinformatically deconvolved and grouped into individual parasites for analysis using modified bcl2fastq and custom python scripts (**Materials and methods**). A schematic to illustrate our experimental workflow is shown in **Figure 1**.

To ensure that our workflow efficiently captures single *Toxoplasma* parasites, we mixed equal numbers of two transgenic lines of RH, one expressing GFP and the other expressing mCherry, and sorted individual parasites into a 384 well plate based on the presence of either green or red signals without a filter for those that were both red and green. After Smart-seq2 amplification, we quantified the expression of GFP and mCherry mRNAs using quantitative polymerase chain reaction (qPCR). Across all 301 wells that we measured, we observed the presence of both GFP and mCherry mRNA in only one well, indicating that the rate of doublet events is below 1% (**Figure 1 - Supplementary Figure 1a**). To address the possibility that the reduced reagent volume in the 384-well format could potentially saturate the reaction chemistry and thus limit quantification range, we sorted varying numbers of RH and quantified with qPCR the mRNA of a gene encoding the abundantly expressed surface protein, SAG1 (**Figure 1 - Supplementary Figure 1b**). The expression values for single, eight, and fifty RH fall in distinct distributions without signs of saturation, indicating that the assay is capable of quantitative measurement at the single *Toxoplasma* level. We then proceeded to sort parasites based on live/dead staining and sequence 729 RH strain single *Toxoplasma* parasites from asynchronous populations grown under tachyzoite conditions. For Pru and ME49 strains, we also collected parasites at several time points post alkaline treatment to follow their expression profiles during *in vitro* transition to bradyzoites (**Materials and methods**), yielding 2655 Pru and 1828 ME49 single parasites. RH reads were aligned to the GT1 strain genome, which is the most complete reference for

Type I parasites, while Pru and ME49 were aligned to the ME49 strain Type II genome reference. Because many genes encoding *Toxoplasma* secretion factors and surface proteins are evolutionary products of gene duplication events[35], we expected high sequence similarity amongst a substantial portion of the parasite genes. Thus, we modified our gene counting pipeline to account for duplicated genes (**Materials and methods**). A comparison of counting methods does not reveal significant differences in the observed counts (**Figure 1 - Supplementary Figure 1c**). Further analysis reveals that our modified pipeline recovered the detection of a few more parasite genes than default parameters (**Figure 1 - Supplementary Figure 1d**).

To ensure that poorly amplified or sequenced parasites did not confound our downstream analysis, we filtered samples based on several quality metrics including percent reads mapping to ERCC spike-in sequences, number of genes detected, and sequencing depth (**Materials and methods**; **Figure 1 - Supplementary Figure 2a**). On average, each sequenced parasite contains 30-50% reads that mapped to *Toxoplasma* genes encoding proteins (top panel in **Figure 1 - Supplementary Figure 2b**). The majority of the unmapped reads are from *Toxoplasma*'s 28s ribosomal RNA. The relatively high rate of rRNA contamination was also observed in single-parasite RNA sequencing of *Plasmodium*[30]. We suspect this occurred due to non-specific priming in the low mRNA content environment of protozoan cells. We normalized for sequencing depth across cells by dividing each read count by the median of read sum to yield "count per median" (CPM). After filtering ERCC spike-in and rRNA genes, we detected on average 996, 1247, and 1067 genes per parasite with greater than 2 CPM (**Materials and methods**) in the RH, Pru, and ME49 datasets, respectively (bottom panel in **Figure 1 - Supplementary Figure 2b**). Characterization of our measurement sensitivity based on logistic regression modeling of ERCC spike-in standards (**Materials and methods**)[36] reveals a 50% detection rate of 10.5, 11.5, and 21.1 molecules for RH, Pru, and ME49 datasets, respectively (top panels in **Figure 1 - Supplementary Figure 2c**). The sensitivity of our 384-well Smart-seq2 measurement is comparable to the previously reported range for the 96-well format[34]. As expected from our qPCR titration experiment, scRNA-seq measurement of gene expression is quantitative at single parasite resolution based on ERCC standards. We determined that the linear dynamic range of our scRNA-seq measurement spans over three orders of magnitude (bottom panels in **Figure 1 - Supplementary Figure 2c**). Taken together, we demonstrate a scalable and cost-effective approach to measure the transcriptomic changes of individual parasites with high sensitivity and accuracy.

Cell cycle landscape of asynchronous Toxoplasma

Previous work posited a potential link between bradyzoite development and cell cycle, which poses a significant challenge to the bioinformatic analysis of either process[21]. To characterize cell cycling changes without confounding contributions from

developmental processes, we first analyzed an asynchronous population of Type I RH strain parasites grown under tachyzoite conditions; this extensively passaged lab strain is known to have little propensity to switch to bradyzoites under such conditions[6] (**Materials and methods**). After filtering out genes whose expression levels did not vary significantly between individual parasites, we projected the data with principal component analysis (PCA) (**Materials and methods**). Interestingly, the first two principal components (PCs) reveal a circular trajectory that coincides with relative DNA content, determined using a cell permeable DNA content stain (top panel in **Figure 2a**). Unsupervised neighborhood clustering identified five distinct clusters of parasites based on their transcriptional profiles (middle panel in **Figure 2a**) (**Materials and methods**). To infer transcriptional dynamics, we applied stochastic RNA velocity algorithm that relies on the ratio of incompletely spliced transcripts to their fully spliced form in order to assess the directionality of transcriptional changes[37,38]. The vector field of RNA velocity indicates a net "counter-clockwise" flow of transcriptional changes (bottom panel in **Figure 2a**) (**Materials and methods**). We assigned cell cycle phase to the clusters based primarily on change in DNA content (**Figure 2 - Supplementary Figure 3a**) but also considering previous bulk transcriptomic characterization[28] (**Figure 2 - Supplementary Figure 3b**). Unsupervised clustering identified two distinct clusters in G_1 state, which we have designated as G_1a and G_1b. We found a list of differentially expressed genes between the two G_1 clusters. The G_1a cluster is highly enriched for the expression of metabolic genes such as phenylalanine hydroxylase (*TGGT1_411100*) and pyrroline-5-carboxylate reductase (*TGGT1_236070*), as well as invasion-related secreted factors such as MIC2 (*TGGT1_201780*), MIC3 (*TGGT1_319560*), and MIC11 (*TGGT1_204530*). On the other hand, G_1b cluster is enriched for the expression of 3-ketoacyl reductase (*TGGT1_217740*) and cytidine and deoxycytidylate deaminase (*TGGT1_200430*), as well as numerous uncharacterized proteins (**Supplementary table 1**). The relative abundance of G_1a, G_1b, S, M, and C states were determined to be 18%, 32%, 28%, 15%, and 7%, respectively. Without chemical synchronization, the correlation between the scRNA-seq data of asynchronous parasites and previously published bulk transcriptomic measurement suggests strong agreement in cluster assignment and cell cycle state identification (**Figure 2 - Supplementary Figure 3b**). This highlights a key advantage of scRNA-seq, as it enables identification of cell cycle status of a parasite without reliance on chemical induction, which may lead to unnatural cellular behavior.

To verify the cyclical nature of gene expression through the lytic cycle, we reconstructed a biological pseudotime of RH using Monocle 2 (**Materials and methods**). The results shown a clear oscillatory expression pattern for the variably expressed genes along the pseudotime axis (**Figure 2b**). To further characterize cell cycle expression patterns, we clustered genes based on pseudotime interpolation and hierarchical clustering (**Materials and methods)**. Some of the key organelles in

tachyzoites are known to be made at different times in the cell cycle[28]. To confirm and refine this finding, we calculated the mean expression values for each set of organelle-specific genes based on their annotation in ToxoDB (**Supplementary Table 2**). This showed the expected, strong oscillation with pseudotime (bottom panel in **Figure 2 - Supplementary Figure 3c**), which also strongly correlates with the oscillation of DNA and total mRNA content (top panels in **Figure 2 - Supplementary Figure 3c**). On the other hand, we also observed instances where a given gene's expression was discordant to the dominant trend of its nominal organelle set (arrows in **Figure 2 - Supplementary Figure 3d**). For example, 63.5% of genes annotated as rhoptry (ROP) or rhoptry neck (RON) are assigned pseudotime cluster 3, while the remaining 36.5% rhoptry genes are assigned pseudotime clusters 1 or 2 (**Figure 2 - Supplementary Figure 3e**). Specifically, genes annotated as ROP33 and ROP34, based on their homology to genes encoding known rhoptry proteins, are assigned to cluster 2 instead of cluster 3 (left panel in **Figure 2 - Supplementary Figure 3f**). Recent reports have experimentally determined these two to be non rhoptry-localizing proteins, thus explaining their discordance[39]. Through analysis of pseudotime clustering, we also identified genes not annotated as ROPs within the ROP-dominated cluster 3, such as *TGGT1_218270* and *TGGT1_230350*, that have recently been shown to encode *bona fide* rhoptry and rhoptry neck proteins, now designated as ROP48 and ROP11, respectively (left panel in **Figure 2 - Supplementary Figure 3f**). As another example, IMC2a peaks in expression level in G_1, while the majority of inner-membrane complex (IMC) genes are expressed towards the M/C phase of the cell cycle (right panel in **Figure 2 - Supplementary Figure 3f**). A recent report has proposed reannotation of IMC2a as a dense granule (GRA) protein (GRA44) based on subcellular localization[40], which is consistent with our unsupervised group assignment of IMC2a as falling in the same cluster 1 where GRA genes dominate. A list of 8590 RH genes with their corresponding pseudotime clustering assignment is provided (**Supplementary Table 3**). We observe high discordance of pseudotime expression for several genes in each annotated organelle sets, suggesting that the current *Toxoplasma* annotation may need significant revision. Our scRNA-seq data provide an important resource to help identify mis-annotated genes and infer putative functions of uncharacterized proteins.

Hidden heterogeneity in asexually developing Toxoplasma

Toxoplasma has one of the most complicated developmental programs of any single-celled organism; however, it is unknown how synchronized the transition is between developmental states. To address this, we assessed the inherent heterogeneity within asexually developing Pru, a type II strain that is capable of forming tissue cysts *in vitro* upon growth in alkaline conditions[18,41]. We applied scRNA-seq to measure and analyze Pru parasites grown in HFFs as tachyzoites ("uninduced") and after inducing the switch to bradyzoites by growth in alkaline media for 3, 5, and 7 days.

Projection of the first two PCs of uninduced Pru tachyzoites (Day 0) reveals the expected circular projection (**Figure 3 - Supplementary Figure 4a**) presumably reflecting cell cycle progression as seen for the RH tachyzoites, described above. To validate this, we developed a random forest classifier model based on our cell cycle assignment in RH (**Materials and methods**). Comparable to what we observed in RH, cell cycle prediction reveals that the uninduced population of Pru is composed of 28%, 41%, 21%, 7%, and 3% parasites in G_1a, G_1b, S, M, and C states, respectively. Consistent with previous observation[23], our data show most induced Pru parasites (Day 3 - 7) are in the G_1 state with a predominance of G_1b (**Figure 3 - Supplementary Figure 4b**).

To identify transcriptomic changes associated with the tachyzoite-bradyzoite transition, we next projected data from both induced and uninduced Pru parasites onto two dimensions using UMAP, a nonlinear dimensionality reduction method (**Materials and methods**)[42]. Unsupervised clustering revealed six distinct clusters of parasites, which we label P1-6 (**Figure 3a**). Cluster formations partially correlate with treatment time points and cell cycle states (**Figure 3b; Figure 3 - Supplementary Figure 4c**), suggesting that the asexual differentiation program may overlap with cell cycle regulation in *Toxoplasma,* as proposed previously[21]. We stratified the datasets by days post alkaline induction (dpi) and observed elevated expression of all bradyzoite marker genes including *SRS44* (*CST1*) and *BAG1* with a concomitant reduction in expression of *SRS29B (SAG1)*, a tachyzoite-specific surface marker gene (**Figure 3 - Supplementary Figure 5**). The abundance of SAG1$^+$ parasites (72%) in the induced population suggests that depletion of this mRNA may be relatively slow and we are measuring SAG1 transcripts made when the parasites were still tachyzoites or that the asexual transition induced by alkaline treatment is highly asynchronous. Interestingly, RNA velocity analysis suggests that P3 may be a fate decision point as the trajectory trifurcates into either P4 (cell cycle), P1, or P6 as evident by the net transcriptional flow (compare **Figure 3a** to right panel in **Figure 3b**).

To determine the gene modules specific to a given cluster, we conducted differential gene expression for each cluster (**Figure 3c** and **Supplementary Table 4**). P1 cluster is correlated with the expression of bradyzoite-specific genes while P2-5 are correlated with the expression of tachyzoite-specific or cell cycle-associated genes (**Figure 3d**). In our scRNA-seq data, we also observe a small portion of *BAG1*$^+$ bradyzoites (7.1%) annotated as either S, M, or C states, indicating that they are replicating (**Figure 3 - Supplementary Figure 4d**). Our data supports the notion that bradyzoites can undergo cell cycle progression, as posited by a previous report[19]. Interestingly, we observe a group of AP2 transcription factors that are differentially expressed across different clusters, some of which are implicated in *Toxoplasma* development (**Figure 3e**). In particular, we identify AP2Ib-1, AP2IX-1, AP2IX-6, and AP2VI-2 as over-expressed in P1, suggesting their potential roles in the regulation of

bradyzoite transition, while AP2-domain protein (TGME49_215895), AP2IX-9, AP2X-8, AP2VIIa-6, AP2XI-1, AP2IX-3, and AP2VIII-7 are highly expressed in P6, hinting at their possible roles in defining this distinct cluster of parasites.

The most highly expressed genes in P6 include genes enriched in P2 as well as bradyzoite-specific genes found in P1 (**Figure 3c**). To identify genes that are specifically expressed in P6, we used Wilcoxon's test (**Figure 3 - Supplementary Figure 6a**) (**Materials and methods)** between P6 and P2 or P1. Comparison of our data to previous bulk transcriptomic measurement in tachyzoites, tissue cyst, or isolates at the beginning or the end of sexual cycle showed no specific enrichment in known developmental stages (**Figure 3 - Supplementary Figure 6b**)[43]. Instead, we show that based on their expression, P6 forms a distinct sub-population of parasites which suggests that alkaline induced *Toxoplasma* may be more heterogeneous than previously thought. Thus, scRNA-seq resolves a transcriptomic landscape of asexual development and suggests the existence of otherwise hidden states.

To determine the reproducibility of the phenomena we observed in the differentiating Pru strain parasites, we repeated the analysis with another widely used Type II strain, ME49, examining 1828 single ME49 parasites exposed to alkaline conditions to induce switching to bradyzoites. Data from the two experiments were computationally aligned using Scanorama to remove technical batch effects while retaining sample-specific differences[44]. Unsupervised clustering revealed 5 distinct clusters in ME49 which share significant overlap in expression patterns with Pru (**Figure 3 - Supplementary Figure 7a**). Matrix correlation of batch-corrected expression across the two strains demonstrate analogous mapping for most, but not all cluster identities (**Figure 3 - Supplementary Figure 7b**). To simplify the visualization and comparison across the two datasets, we next applied Partition-Based Graph Abstraction (PAGA) to present clusters of cells as nodes with connectivity based on similarity of the transcriptional profiles between clusters[45]. In particular, a side-by-side comparison of expression of tachyzoite, bradyzoite, and sexual stage specific genes reveals some key similarities and dissimilarities (**Figure 3 - Supplementary Figure 7c**). Clusters P1 and M1 are both enriched for the expression of bradyzoite marker genes. Clusters M4-5 and P4-5 are both predicted to be S/M/C phases of the cell cycle. While most ME49 clusters express tachyzoite marker genes, enolase-2 and LDH-1, which have previously been described as relatively tachyzoite-specific[46], are expressed at much lower level than in Pru. Curiously, P6-specific genes (green panels in **Figure 3 - Supplementary Figure 7c**) are not enriched in any cluster in ME49, suggesting a lack of corresponding P6 cluster in ME49. Such differences may not be surprising, however, as Pru and ME49 have entirely distinct passage histories, although both were grown in our laboratory exclusively *in vitro* as tachyzoites on HFFs over at least 2 years prior to the experiments described here. Because measurement sensitivity in ME49 (21 molecules) was lower than that of Pru (11 molecules) which reduces the ability to differentiate technical

dropouts from differentially expressed genes, we focused further analysis on the Pru dataset.

Transcriptional variation between parasites independent of measurement noise, cell cycle, or asexual development

A unique advantage of scRNA-seq over bulk RNA-seq is its ability to measure cell-to-cell variation that is stochastic in nature or independent of known biological processes. We have developed a computational approach to identify genes with such variation. While our scRNA-seq is sensitive, one needs to measure the level of noise in order to determine true, intrinsic variability of parasite expression program. Noise levels can be estimated using the ERCC synthetic RNA spike-ins that were added in differing amounts to each sample and then fitting a logistic regression to model the expected detection rate, as shown in **Figure 1 - Supplementary Figure 2c**. This allowed us to determine whether the expression level of a given variable gene in *Toxoplasma* is above the detection limit and thus whether its variation is readily explained by measurement noise. Next, for the variable genes that are above detection limit, we asked if their variability can be explained by either cell cycle or developmental state, the two biological variables that, as expected, show a major influence on gene expression in our system. To do this, we perform “projection scoring”, in which we use a bootstrapped K-nearest neighbor (KNN) approach that quantifies the dependence of a gene’s expression variability on the PCA and UMAP projections. Genes that vary as a result of cell cycle or development are expected to show similar expression levels in neighboring cells in the projection and different expression levels in cells that are widely separated (**Figure 4a**).

Applying this approach to our Pru data set, we first see that, as expected for an asynchronous population at different cell cycle and developmental states, mRNA for many genes has a low detection rate even though those genes have a mean abundance across all cells that is above our threshold for detection (**Figure 4b**) (**Materials and methods**). Comparison between RH and Pru demonstrates congruence of many “variant” genes: 213 shared genes are more variable than the ERCC spike-ins in both datasets (**Figure 4 - Supplementary Figure 8a**). Some degree of disagreement between the two datasets is expected as the Pru data include differentiating parasites while the RH data do not. Starting from the list of “variant” genes that we identified in Pru, projection-scoring quantified the dependence of each gene on the PCA embedding, reflecting cell cycle progression, and UMAP embedding, reflecting asexual development and cell cycle progression. Comparison of projection scores between RH and Pru shows consistency in cell cycle dependence, revealing that while the variance of some genes (e.g., ROP genes) is readily explained by cell cycle, a large fraction of variable genes shows no correlation with cell cycle (upper right and lower left areas of top panel in **Figure 4c**, respectively). As expected from our Pru data, gene dependence

on PCA and UMAP projections are highly correlated (bottom panel in **Figure 4c**); however, projection scoring identifies a subset of genes whose variation depends exclusively on asexual development, but not on cell cycle, including ones that we identified previously as enriched in bradyzoites (**Supplementary Table 4**). This shows that projection scoring can be used to discover genes that may differ in regulation across different dimensions of intrinsic biological variability.

To determine the variation dependence on cell cycle and asexual development in Pru, we quantified projection scores for organelle gene sets of “variant” genes and ERCC spike-ins (top panel in **Figure 4d**). The results show a wide distribution of dependence across different organelle sets in the biological data. As expected, ERCC spike-ins, which are randomly distributed between samples, exhibited low projection scores on both cell cycle and asexual development projections. We took the upper end of ERCC score (~0.35) as a threshold to further classify each variant gene. For variant genes with scores above the threshold in either asexual development or cell cycle dependence, they are classified as “Dependent”, otherwise they are considered “Independent” of either process. Expression variability of most rhoptry (ROP), microneme (MIC), and inner-membrane complex (IMC) proteins show high dependence on cell cycle and/or asexual development (bottom panel in **Figure 4d**). On the other hand, >40% of SRS surface antigens, GRAs, and other non-parasite-specific or unannotated genes have low dependence score on both of these two biological processes. We show several SRS surface antigens as examples of genes whose expression shows low cell cycle and asexual development dependence in Pru data (**Figure 4 - Supplementary 8b**), highlighting the variation of their expression between neighboring cells on the projection. Thus, our analysis reveals that a substantial fraction of variable genes contributes to previously undetermined parasite-to-parasite variation that would not be detectable in bulk transcriptomic analyses.

Discussion

We describe here single-cell RNA sequencing (scRNA-seq) for measurement of mRNA transcripts from individual *Toxoplasma gondii*, an obligate intracellular protozoan parasite. The results show that scRNA-seq can reveal intrinsic biological variation within an asynchronous population of parasites. Two types of biological variation could be seen in our asynchronous populations: cell cycle progression and asexual differentiation. We found the existence of two distinct 1N transcriptional states in cycling parasites which we call G_1a and G_1b, concurring with what was previously reported in *Toxoplasma*[28]. Interestingly, bradyzoites are found predominantly in G_1b but not in G_1a state, suggesting the possibility of a putative checkpoint between these two phases that may also play a role in regulating the developmental transition. Our data further shows a small fraction of bradyzoites to be cycling which supports the hypothesis that bradyzoites can in fact divide[22]. Our results showed a very strong correlation between cell cycle and expression of genes encoding proteins in various subcellular organelles, as noted previously using synchronized bulk populations[28]. The results here, however, show an even more dramatic and extreme dependence on cell cycle, allowing refinement of approaches that use such timing to predict a given protein's ultimate organellar destination in the cell[47]. They also extend such analyses to the Type II strains, Pru and ME49, which have not previously been examined in this way.

In addition to the above, we observed some striking and unexpected heterogeneity within asexually developing parasites. We discovered a cluster of cells, labeled P6, in the differentiating Pru parasites that is distinct from the rest of the alkaline-induced population of cells. Constituting 21% of the alkaline-induced population, the P6 cluster is marked by a set of genes that were previously detected by bulk transcriptomics in bradyzoites of tissue cysts[43]. Remarkably, while most of these genes have unknown functions, we identified an enriched gene with predicted AP2 domain, which may contribute to the unique expression pattern observed in this group of parasites. We found that P6 expression profile is intermediate to P2 tachyzoites and P1 bradyzoite clusters. Interestingly, the genes enriched in P6 overlap with a subset of canonical bradyzoite marker genes including *LDH2* and *SRS35A*, albeit expressed at a lower level than in P1 (**Figure 3c**). In addition, we observed a gradual increase in the proportion of P6 cells as induction proceeded from day 3 to day 7. Taken together, one possible explanation for the emergence of P6 cluster is a reverted conversion from bradyzoites to tachyzoites in which alkaline stress fails to maintain the bradyzoite state. Our data and previous reports are consistent with this interpretation[48]. On the other hand, we cannot rule out the possibility that this cluster is developmentally "confused" by the presence of a general stressor such as alkaline. RNA velocity analysis in the Pru data does not reveal a strong transcriptional flow between P1 and P6. Rather, P6 appears to transcriptionally transition from P2 and P3 tachyzoites. Thus, the P1 bradyzoites and P6 parasites are either distinct and separate developmental

trajectories, or the transition from P1 to P6 is a rapid and rare event. Regardless, our results reflect a surprising diversity in an asexually transitioning population of *Toxoplasma*. Future measurement of single parasites isolated from *in vivo* sources coupled with genetic manipulation of the parasite genome, will further clarify the causality and relevance of developmental states that we identified here.

To address how much cell-to-cell variability there is between parasites of similar developmental states, we developed a novel approach based on random permutation and K-nearest neighbor (KNN) averaging to quantify the association of expression variation to known biological processes, like cell cycle and development, that underlie PCA and UMAP projections of scRNA-seq. Combined with the analysis of ERCC synthetic spike-ins, this allowed us to tease out expression variation in single parasites that results from one of the biological processes as well as measurement noise. Previous reports have noted potential issues with ERCC spike-ins in estimating technical variations of endogenous mRNAs, potentially due to differences in poly-A tail lengths and the lack of 5' cap[34,49,50]. Our results show that many low abundant endogenous parasite genes have significantly higher detection rate than would be predicted by ERCC with similar abundance, suggesting that ERCC spike-ins provide, as previously reported, a conservative underestimate of the detection sensitivity of endogenous genes[50]. Intriguingly, the resulting analysis showed that this single-celled organism exhibits unexplained variation in the expression of several genes. Whether such a pattern of variation may define novel cellular subtypes will require further experimentation to probe the stability and stochasticity of the expression of these genes. Interestingly, amongst these projection-independent genes, we found the expression of several SRS surface antigen genes, which are known to play a role in host attachment, and dense granule genes, which are known to play a role in intracellular interaction with host, to be highly variable between cells of similar developmental states. We also find other non-parasite-specific genes, including genes encoding metabolic enzymes, to be highly projection-independent. While we cannot exclude the possibility that variation in these cells is due to stochastic bursts of transcription from these genes, especially given the small size of *Toxoplasma*, it is possible that such variability has biological meaning. For example, it could expand the mode of interactions with the host and be the result of strong selective pressure to maximize invasion efficiency and transmission in a variety of different host species of cell types. Maintaining a diverse phenotypic diversity can be beneficial in ensuring at least some members will be able to propagate in whatever the host environment encountered. The biological implication of single-celled parasite variation and its relevance to *in vivo* infection will be an important area of investigation for future studies. We see the application of single-cell co-transcriptomic sequencing of both the host cell and the parasite as a potentially powerful approach to further deconstruct the complexity of parasite-host interactions.

Materials and Methods

Cell and Parasite Culture

All *Toxoplasma gondii* strains were maintained by serial passage in human foreskin fibroblasts (HFFs) cultured at 37 C in 5% CO_2 in complete Dulbeco's Modified Eagle Medium (cDMEM) supplemented with 10% heat-inactivated fetal bovine serum (FBS), 2 mM L-glutamine, 100 U/ml penicillin, and 100 ug/ml streptomycin. *T. gondii* strains used in this study were RH, Pru GFP[12], and ME49-GFP-luc[51].

In vitro Bradyzoite Switch Protocol

Differentiation to bradyzoite was induced by growth under low-serum, alkaline conditions in ambient (low) CO_2 as previously described[53]. Briefly, confluent monolayers of HFFs were infected with tachyzoites at a multiplicity of infection (MOI) of 0.025 in RPMI 1640 medium (Invitrogen) lacking sodium bicarbonate and with 1% FBS, 10 mg/ml HEPES, 100 U/ml penicillin, and 100 g/ml streptomycin at pH 8.2. The infected HFFs were cultured at 37°C without supplemented CO_2.

Preparation of Parasites for Fluorescence Activated Cell Sorting (FACS)

HFF monolayers infected with parasites overnight were scraped, and the detached host cells were lysed by passing them through a 25-gauge needle three times or a 27-gauge needle six times. The released parasites were spun down at 800 rpm for 5 minutes to pellet out host cell debris, and the supernatant was spun down at 1500 rpm for 5 minutes to pellet the parasites. The parasites were then resuspended in 500 µL of FACS buffer (1x PBS supplemented with 2% FBS, 50 ug/ml DNAse I, and 5 mM $MgCl_2*6H_2O$), passed through both a 5 µm filter and a filter cap into FACS tubes, and stored on wet ice until it was time to sort. In samples stained for DNA content, the parasites were resuspended in 500 µL of FACS buffer plus 1.5 µL of Vybrant DyeCycle Violet (from ThermoFisher, catalog number V35003) and incubated at 37 C and 5% CO_2 for 30 minutes.

The parasites were also stained with either propidium iodide (PI), Sytox Green, or the live/dead fixable blue dead cell stain kit (catalog number L34962) prior to sorting in order to distinguish live cells from dead cells. To stain with PI, 10 µL of 0.5 mg/ml PI was added to every 500 µL of parasite suspension in FACS buffer, and the parasites were incubated covered on ice for at least 15 minutes. To stain with Sytox Green, 1 drop of Sytox Green per ml was added to the parasite suspension in FACS buffer, and the parasites were incubated at room temperature for at least 15 minutes. To stain with the live/dead fixable blue dead cell stain kit, 1.5 µL of the kit's viability dye was added to every 500 µL of parasites along with the secondary antibody, and parasites were washed and resuspended in FACS buffer as usual.

FACS of parasites

Eight mL of lysis buffer was prepared by mixing together: 5.888 mL of water, 160 µL recombinant RNase inhibitor (Takara Clonetech), 1.6 mL of 10 mM dNTP (ThermoFisher), 160 µL of 100 uM oligo-dT (iDT; see attached "supplementary_file1_oligos.csv" for oligos), 1:600,000 diluted ERCC spike-in (ThermoFisher), and 32 µL of 10% Triton X-100. All reagents are declared RNase free. Lysis plates were prepared by dispensing 0.4 µL of lysis buffer into each well of a 384 well hard-shell low profile PCR plate (Bio-rad) using liquid handler Mantis (Formulatrix). Single parasites were sorted using the Stanford FACS Facility's SONY SH800s sorter or BD Influx Special Order sorter into the 384-well plates loaded with lysis buffer. Single color and colorless controls were used for compensation and adjustment of channel voltages. The data were collected with FACSDiva software and analyzed with FlowJo software. RH parasites were index sorted with fluorescence signal of cell permeable DNA stain, DyeCycle Violet.

Single-Toxoplasma cDNA synthesis, library preparation, and sequencing

Smart-seq2 protocol was carried out as previously described[54] using liquid handlers Mantis and Mosquito (TTP Labtech) using a 2 µL total volume. We conducted 19 rounds of cDNA pre-amplification after reverse transcription. Each well is then diluted with 1 to 4 v:v in RNAse free elution buffer (QIAgen) to a total volume of 8 µL. Then, we conducted library preparation with in-house Tn5 tagmentation using custom cell barcode and submitted for 2 x 150 bp paired-end sequencing on NovaSeq 6000 at the Chan Zuckerberg Biohub Genomics core.

Quantitative polymerase chain reaction (qPCR) for parasite benchmark

To quantify the purity of single parasite sort and to ensure the cDNA synthesis reaction was not saturated, GFP, mCherry, or SAG1 mRNA expression were measured using commercial qPCR mastermix, SsoAdvanced™ Universal SYBR Green mastermix (Bio-rad). Briefly, 0.1 µL of diluted cDNA was added in a total of 2.1 µL reaction volume per well on 384 well plate with qPCR mastermix and 200 nM PCR primers. The reaction was incubated on a Bio-rad qPCR thermal cycler with the following programs: 5 minutes of 95ºC, 45 cycles of 95ºC for 5 seconds, 56ºC for 1 minute, and imaging. The primer sequences are provided in "supplementary_file1_oligos.csv".

Sequencing alignment

BCL output files from sequencing were converted into gzip compressed FastQs via a modified bcl2fastq demultiplexer which is designed to handle the higher throughput per sequencing run. To generate genome references with spike-in sequences, we concatenated ME49 or RH genome references (version 36 on ToxoDB) with ERCC sequences. The raw fastq files are aligned to the concatenated genomes

with STAR aligner (version 2.6.0c) using the following settings: "--readFilesCommand zcat --outFilterType BySJout --outFilterMutlimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --outSAMstrandField intronMotif --outSAMtype BAM Unsorted --outSAMattributes NH HI AS NM MD --outFilterMatchNminOverLread 0.4 --outFilterScoreMinOverLread 0.4 --clip3pAdapterSeq CTGTCTCTTATACACATCT --outReadsUnmapped Fastx". Transcripts were counted with a custom htseq-count script (version 0.10.0, https://github.com/simon-anders/htseq) using ME49 or RH GFF3 annotations (version 36 on ToxoDB) concatenated with ERCC annotation. Instead of discarding reads that mapped to multiple locations, we modified htseq-count to add transcript counts divided by the number of genomic locations with equal alignment score, thus rescuing measurement of duplicated genes in the Toxoplasma genome. Parallel jobs of STAR alignment and htseq-count were requested automatically by Bag of Stars (https://github.com/iosonofabio/bag_of_stars) and computed on Stanford high-performance computing cluster Sherlock 2.0. Estimation of reads containing exonic and intronic regions is computed with Velocyto estimation on the BAM output files and requested automatically by Bag of Velocyto (https://github.com/xuesoso/bag_of_velocyto) on Sherlock 2.0. Gene count matrix is obtained by summing up transcripts into genes using a custom python script. Scanpy velocyto package is then used to estimate transcriptional velocity on a given reduced dimension. Parameters used for generating the results are supplied as supplementary python scripts. Sample code to generate the analysis figures are provided in supplementary jupyter notebooks.

Data preprocessing

To filter out cells with poor amplification or sequencing reaction and doublet cells, we discarded cells based on gene counts (>0 reads), total reads sum, percent reads mapped to *Toxoplasma* genome, percent ERCC reads, and percent ribosomal RNA reads. Next, we filtered "ribosomal RNA" genes from the gene count matrix. Gene count matrices are normalized as counts per median (CPM):

$$X_{norm} = \frac{X}{\sum(X)} \cdot median(\sum(X)) \qquad (1)$$

where *X* is the gene count matrix, *sum(X)* is the read sum for each cell, and *median(sum(X))* is the median of read sums. Normalized data are added with a pseudocount of 1 and log transformed (e.g. $\log_2(X_{norm}+1)$). To determine the detection limit (e.g. 50% detection rate), we modeled the detection probability of ERCC standards with a logistic regression as a function of spike-in amount[33].

We calculated an estimate of absolute molecular abundance for all genes by fitting a linear regression to ERCC spike-ins:

$$log_2(Y) = m \cdot log_2(X_{norm} + 1) + b \qquad (2)$$

where $X_{norm,\ ERCC>0.5}$ is the observed CPM value for ERCC spike-ins above the detection limit, *Y* is the amount of ERCC spike-in, *m* is the regression coefficient, and *b* is the intercept. To reduce the influence of measurement noise, we fit the model only to ERCC spike-ins with mean expression above the detection limit.

Cell cycle analysis and annotation

To determine the transcriptional variation associated with cell cycle, we applied Self-Assembling Manifolds (SAM)[55] to filter for highly dispersed gene sets (>0.35 SAM weights) in asynchronous RH population. Principal components analysis (PCA) is then applied to the filtered and normalized RH data, and the nearest neighbor graph (K=50) is computed using "correlation" as a similarity metric. We identified the putative "G1" clusters with 1N based on DNA content stain. Parasites in "G1" cluster are further sub-clustered with Louvain Clustering, in which we identified "G_1a" and "G_1b" clusters with distinct transcriptional profiles. Pearson correlation between single-cell and bulk transcriptomic data is computed between bulk assignment[28] and the scRNA-seq cluster assignment through which each cluster is uniquely assigned with a cell cycle state. To quantify genes that are differentially expressed across cell cycle clusters, we applied Kruskal-Wallis test. Genes are considered differentially expressed if their p-values are less than 0.05 and they are at least 2-fold over-expressed in a cluster compared to the average expression level of other clusters. We computed differential expression across all cell cycle clusters as well as between the "G_1a" and "G_1b" clusters; the results are uploaded as supplementary tables 1 and 2, respectively. To enable cell cycle assignment transfer from RH to Pru and ME49 data, we implemented a random forest classification model trained on RH data. Briefly, this is done by training a model with 1000 estimators on L2-normalized RH expression data containing only cell cycle associated genes in a 60-40 split scheme. Then the model is applied to predict cell cycle labels of L2-normalized Pru or ME49 data containing the homologous cell cycle associated genes. The testing accuracy was over 95%.

Pseudotime construction and clustering

Pseudotime analysis is conducted with Monocle 2 package in R on preprocessed dataset with highly dispersive genes as described previously. A cell in "G_1a" is designated as the root cell, and all other cells are placed after this cell in order of their inferred pseudotime. To cluster genes based on their pseudotime expression pattern, high frequency patterns are removed through a double spline smoothing operation. The

interpolated expression matrix is then normalized by maximum expression along pseudotime such that the maximum value of gene expression along pseudotime is bound by 1. We then applied agglomerative clustering on this interpolated and normalized expression matrix using “correlation affinity” as similarity metric and “average linkage” method to predict three distinct clusters of genes.

Measurement noise analysis and projection scoring

To identify genes with greater variability than can be explained by measurement noise, we first modeled probability of detection as a logistic function of ERCC spike-in mean abundance:

$$P_{detection} = \frac{1}{1+e^{-\beta \cdot log_2(\underline{X}_{norm})+c}} \quad (3)$$

where β and c are parameters of the model, and $\underline{X}$ is the mean abundance for a given ERCC sequence. We then computed Z_i, the z-score of detection deviation from the logistic fit, for each gene:

$$Z_i = \frac{D_i - E(D)}{\sqrt{var(D)}} \quad (4)$$

where D_i is the difference between detection rate of a gene and its predicted detection rate given its mean abundance, *E(D)* and *var(D)* are the expectation values and variance of detection difference for all genes, respectively. *Z* is converted to p-values assuming an one-sided Gaussian distribution of null values. Genes with p-values lower than 0.05 and lower detection probability than the estimated fit are considered variant.

To quantify the dependence of expression variation on a two-dimensional projection, we developed a novel approach based on k-nearest neighbor (KNN) averaging. First, a KNN graph is computed by locating nearest neighborhood in an arbitrary two-dimensional projection using euclidean distance. We then generated a null expression matrix by shuffling the gene expression matrix along each cell column, such that its correlation with respect to the coordinate on projection is completely lost. Next, we compute an updated gene expression value by taking the average of expression values across the KNN. This is equivalent to:

$$X_{KNN} = \frac{M}{k} \cdot X_{norm} \quad (5)$$

where X_{KNN} is the updated KNN averaged expression, *M* is the nearest-neighbor graph with *k* being the number of nearest neighbor, and *X* is the log-transformed CPM of observed or null expression matrices. We chose a *k* of 5 for all our analysis as varying *k* did not have a large effect on the results (data not shown). In our experiments, we have shown that the first two principal components (PCs) of PCA on RH and Pru correspond to the projection projection of cell cycle progression, and a two-dimensional UMAP projection of Pru corresponds to asexual development and cell cycle progression. We thus computed X_{KNN} for both the original, observed expression matrix and the shuffled, null matrix on either projection to reflect dependence on cell cycle progression and/or

asexual development. X_{KNN} is further normalized to have identical sum as the original expression values. A Kolmogorov-Smirnoff two sample test is computed between the normalized X_{KNN} of the observed matrix and that of the shuffled matrix based on 100 random permutations. The projection-dependence score for each gene is then computed as:

$$S_g = \sqrt{-log(\bar{P_g})} \qquad (6)$$

where S_g is the projection-dependence score for gene g and $\bar{P_g}$ is the average p-values of 100 tests. We present S_g normalized by the maximum score within each respective data set.

Acknowledgements

We thank Fabio Zanini, Felix Horns, and Geoff Stanley for illuminating discussion and advice to YX on experiments and analysis. We thank Saroja Korullu, Robert Jones, and Vickie Lin for assistance with library preparation and sample submission. This study is supported by National Institute of Health (NIH) RO1 AI21423, AI29529, and Chan Zuckerberg Biohub. YX and TCT are supported by Stanford Interdisciplinary Graduate Bio-X Fellowships. SR is supported by NIH F30 AI124589-03. AF is supported by NIH 5T32AI007328-30 and a Gilliam Fellowship for Advanced Study from Howard Hughes Medical Institute.

Figures

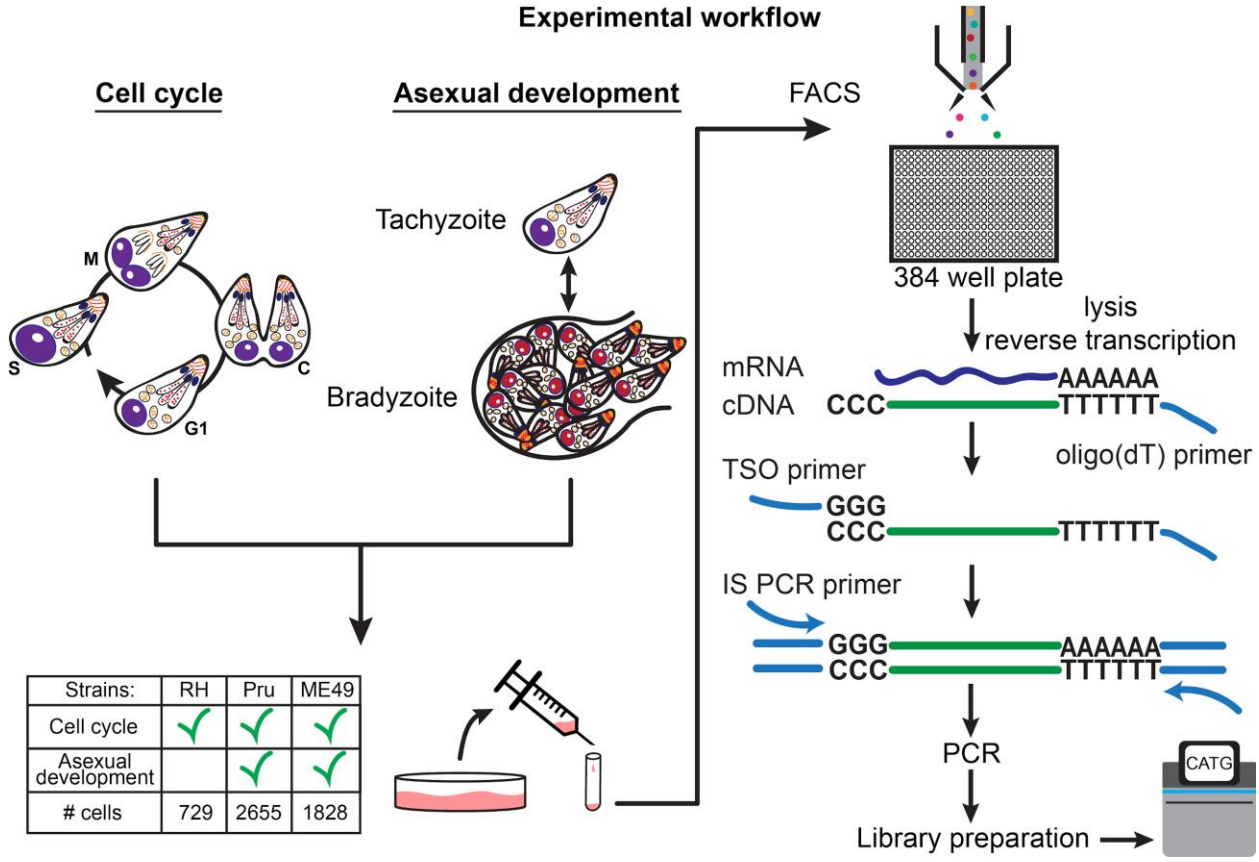

Figure 1. Schematic of single-cell RNA-sequencing (scRNA-seq) based on a modified Smart-seq2 protocol for 384-well plate. A table of strain types with the number of sequenced samples is provided.

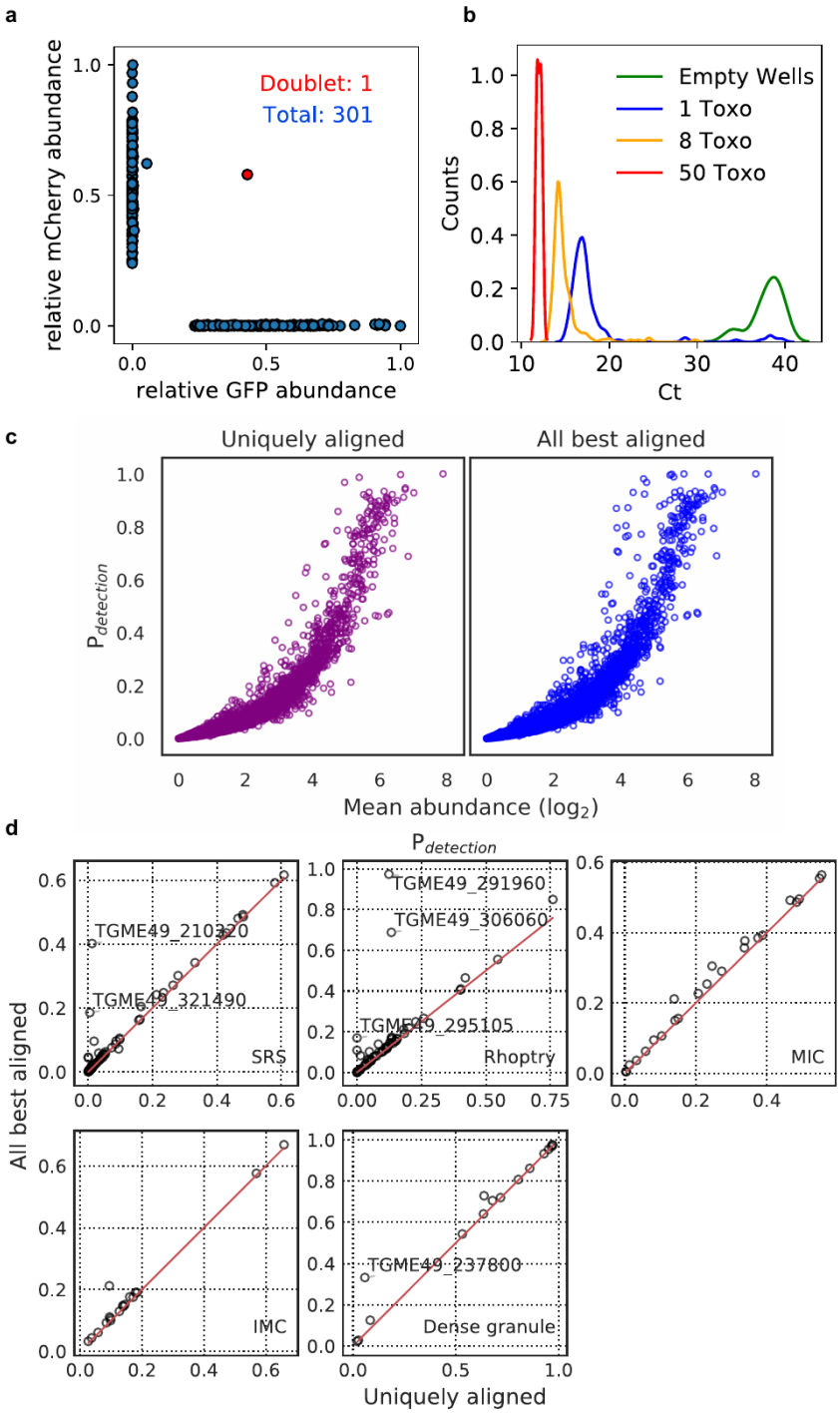

Figure 1 - Supplementary Figure 1. (a) qPCR measurement of mRNA expression in 302 transgenic *Toxoplasma* cells expressing GFP or mCherry mixed at 1:1 ratio. **(b)** qPCR Ct values of abundant surface protein, SAG1, measured for 374 wells with zero, one, eight, or fifty sorted parasites at 16, 176, 176, and 6 replicates, respectively. **(c)** Comparison between "uniquely aligned" (default htseq-count settings) and "all best aligned" (count each feature with equal read alignment score) in the detection rate in Type I strain, RH. **(d)** A more detailed comparison of detection rate of several parasite-specific gene sets. Genes that are detected more frequently in "All best aligned" setting are annotated in the plot.

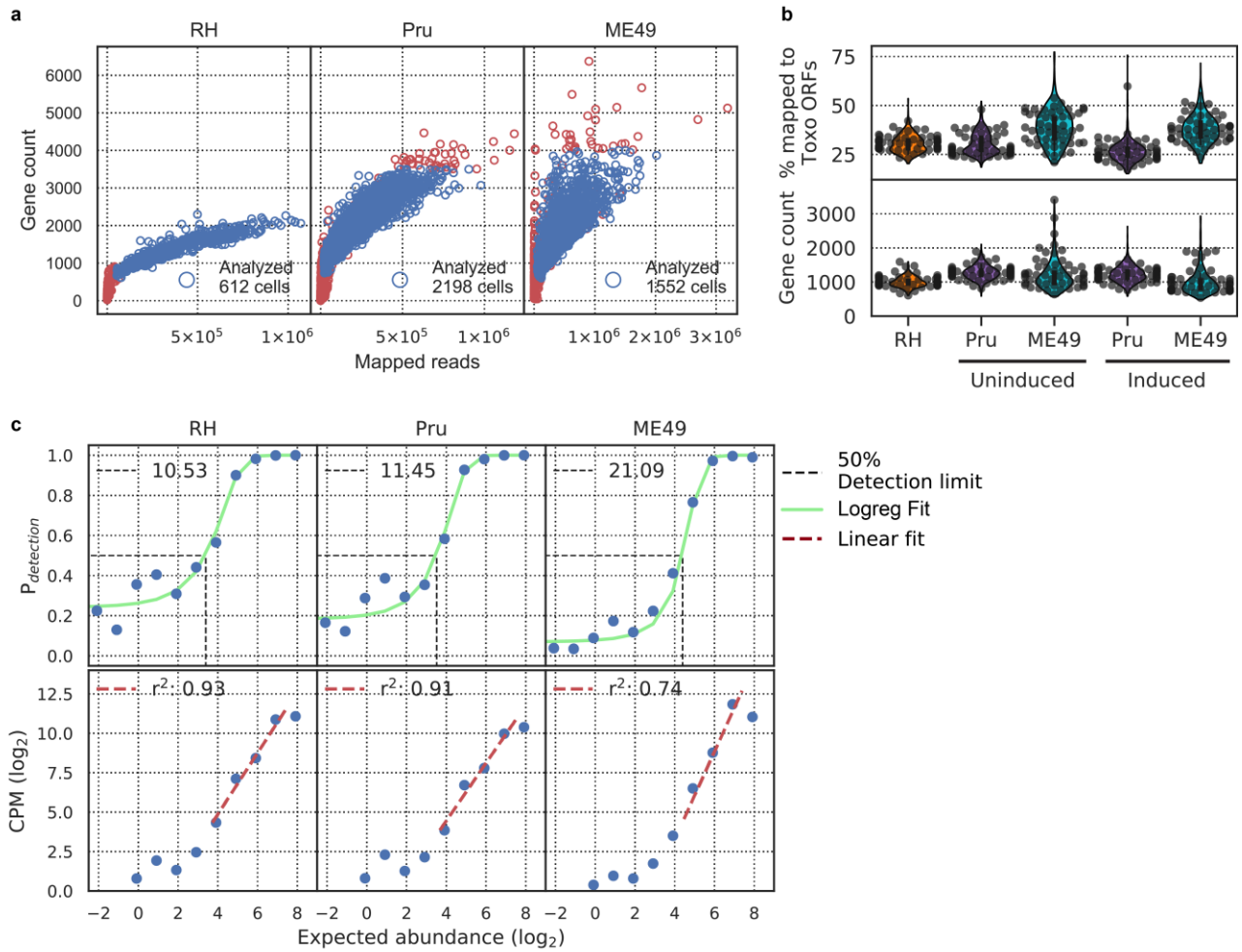

Figure 1 - Supplementary Figure 2. (a) Comparison of gene counts (>2 CPM) and total mapped read counts for RH, Pru, and ME49 from left to right, respectively. Text in lower right corner indicates the number of parasites that passed cell filtering and were analyzed (blue open circles). **(b)** Top panel: distributions of percentage of reads in analyzed cells that mapped to Toxoplasma Open Reading Frames (ORFs). Bottom panel: distributions of gene counts (>2 CPM) in analyzed cells. Uninduced Pru and ME49 were grown in the absence of alkaline (Day 0), whereas induced Pru and ME49 were grown in the presence of alkaline (Day 3 - 7). **(c)** Top panel: Logistic regression modeling (green line) of detection limit (50% detection rate, black dotted line) of ERCC spike-ins. Text on top left of each sub-panel indicates the detection limit in absolute molecular counts. Bottom panel: Linear regression modeling (crimson line) of measurement accuracy fitted on ERCC spike-ins with abundance above the detection limit. Text on top left of each sub-panel indicates the coefficient of determination for the regression fit.

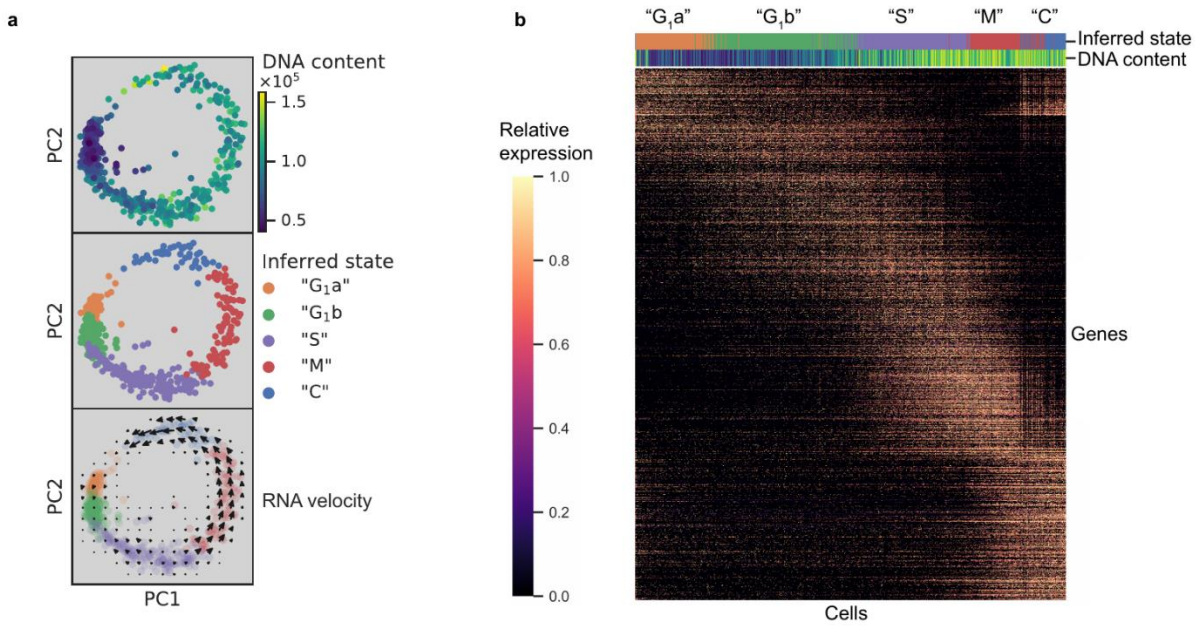

Figure 2. (a) Projection of the first two principal components in RH data set. Top panel: 612 RH cells are colored by fluorescence measurement of a cell permeable DNA content stain. Center panel: cells are colored by cluster assignment and labeled by the inferred “cell cycle” state. Bottom panel: RNA velocity vector field is overlaid on top of the inferred state colors, with arrows pointing in the direction of net transcriptional change. **(b)** Heatmap of the 1465 most variable gene expression ordered by pseudotime assignment from left to right. Top colorbar reflects the assignment of inferred state and bottom colorbar reflects the relative fluorescence of DNA content using the same color scheme as in (a).

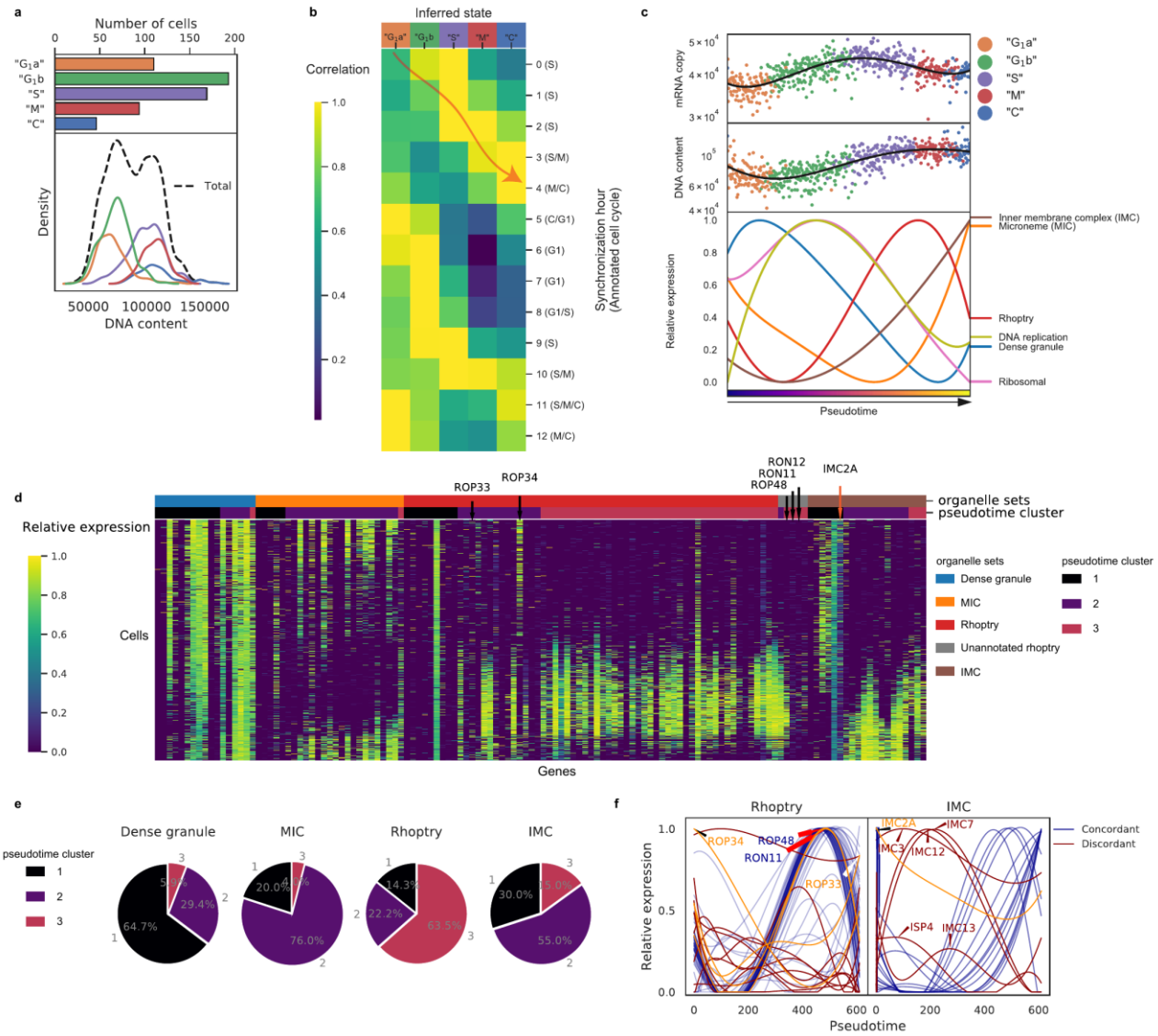

Figure 2 - Supplementary Figure 3. (a) Top panel: Numbers of RH parasites in each inferred "cell cycle" state. Bottom panel: Density plots of DNA content distributions stratified by the inferred state. **(b)** Heatmap of average expression correlation between each inferred "cell cycle" state of RH and each time-point of bulk transcriptomic measurement based on chemically synchronized parasites[28]. **(c)** Absolute mRNA abundance (top panel) and DNA content (center panel) ordered by "cell cycle" pseudotime with individual cells colored by their inferred states. A spline smoothing is applied to approximate a rolling average along the pseudotime (black solid line). Average expression of gene sets based on ToxoDB (v.36) annotation of organellar destination of the protein product after double spline smoothing (bottom panel). **(d)** Heatmap of gene expression ordered by organelle sets (top colorbar) and pseudotime cluster (bottom colorbar). "Unannotated rhoptry" refers to genes not annotated in ToxoDB (v.36) as encoding a rhoptry protein but whose expression pattern is highly concordant with the dominant rhoptry pattern. **(e)** Pie charts of pseudotime cluster frequency for parasite organelle sets. **(f)** Expression of annotated rhoptry (left panel) and inner-membrane complex (IMC; right panel) genes along pseudotime with different colors indicating genes concordant (blue) and discordant (crimson and orange) to the major trend of their organelle sets.

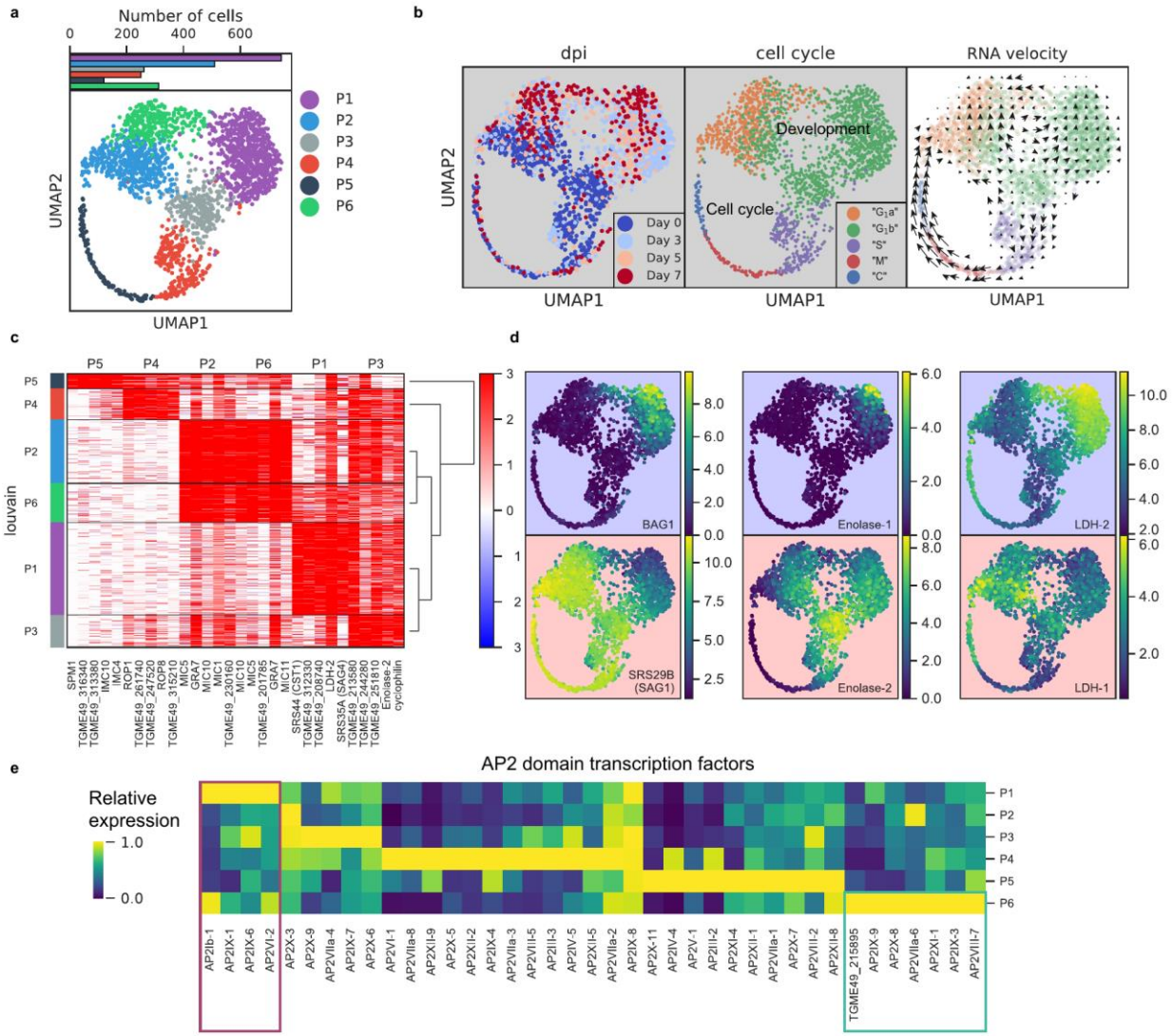

Figure 3. (a) UMAP projection of 809 uninduced and 1389 induced Pru parasites with colors indicating Louvain cluster assignment. Top panel shows the number of parasites in each cluster. **(b)** UMAP projections of Pru parasites colored or labeled by days post induction (dpi), inferred cell cycle states, and RNA velocity from left to right. **(c)** Heatmap of differentially expressed genes (along columns) across Louvain clusters of cells ordered by hierarchical clustering (along rows). The top 5 most enriched genes from each cluster are presented. **(d)** UMAP projections of Pru colored by the neighbor-averaged expression ($\log_2$ CPM) of bradyzoite (top panels, purple background) and tachyzoite (bottom panels, red background) marker genes. **(e)** Heatmap of differentially expressed AP2 transcription factor in Louvain clusters. Purple and green rectangles highlight AP2s enriched in clusters P1 and P6, respectively.

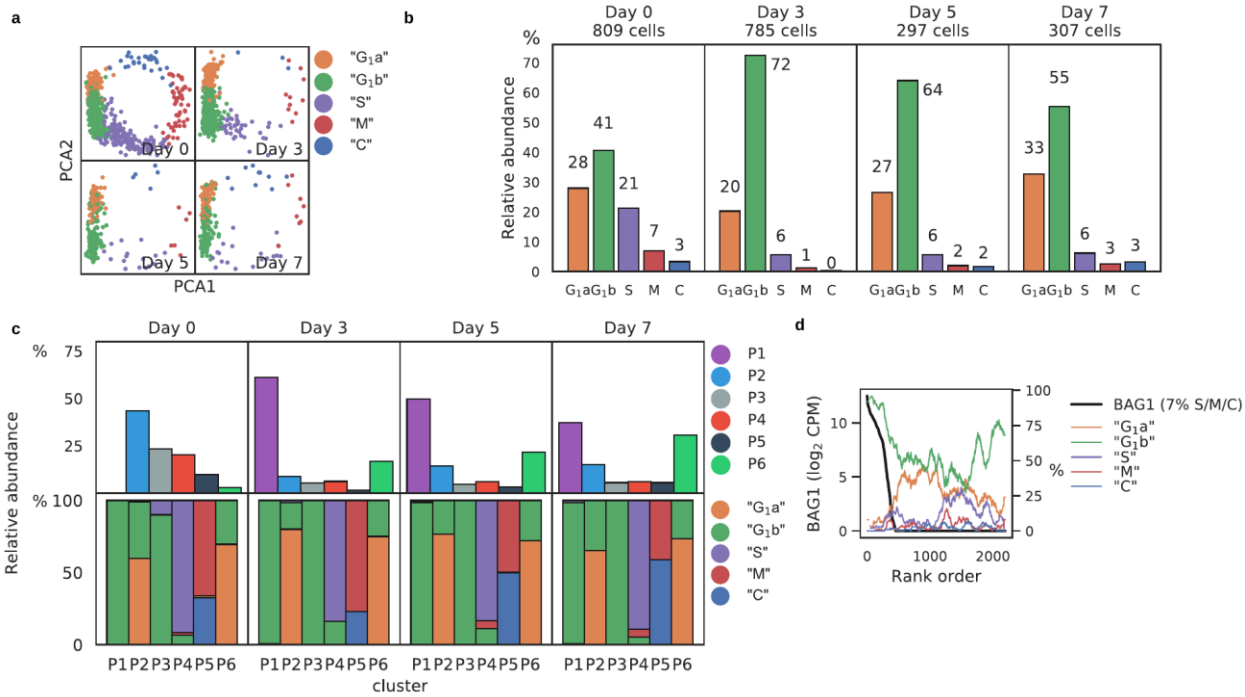

Figure 3 - Supplementary Figure 4. (a) PCA projection of Pru stratified by days post induction (dpi) and colored by predicted cell cycle state. **(b)** Frequency of predicted cell cycle states at different dpi time points. **(c)** Frequency of Louvain clusters (top panels) and predicted cell cycle states in each cluster (bottom panels). **(d)** Rolling average frequency of predicted cell cycle states (colored lines) ordered by expression level of the canonical bradyzoite marker, *BAG1* (black line).

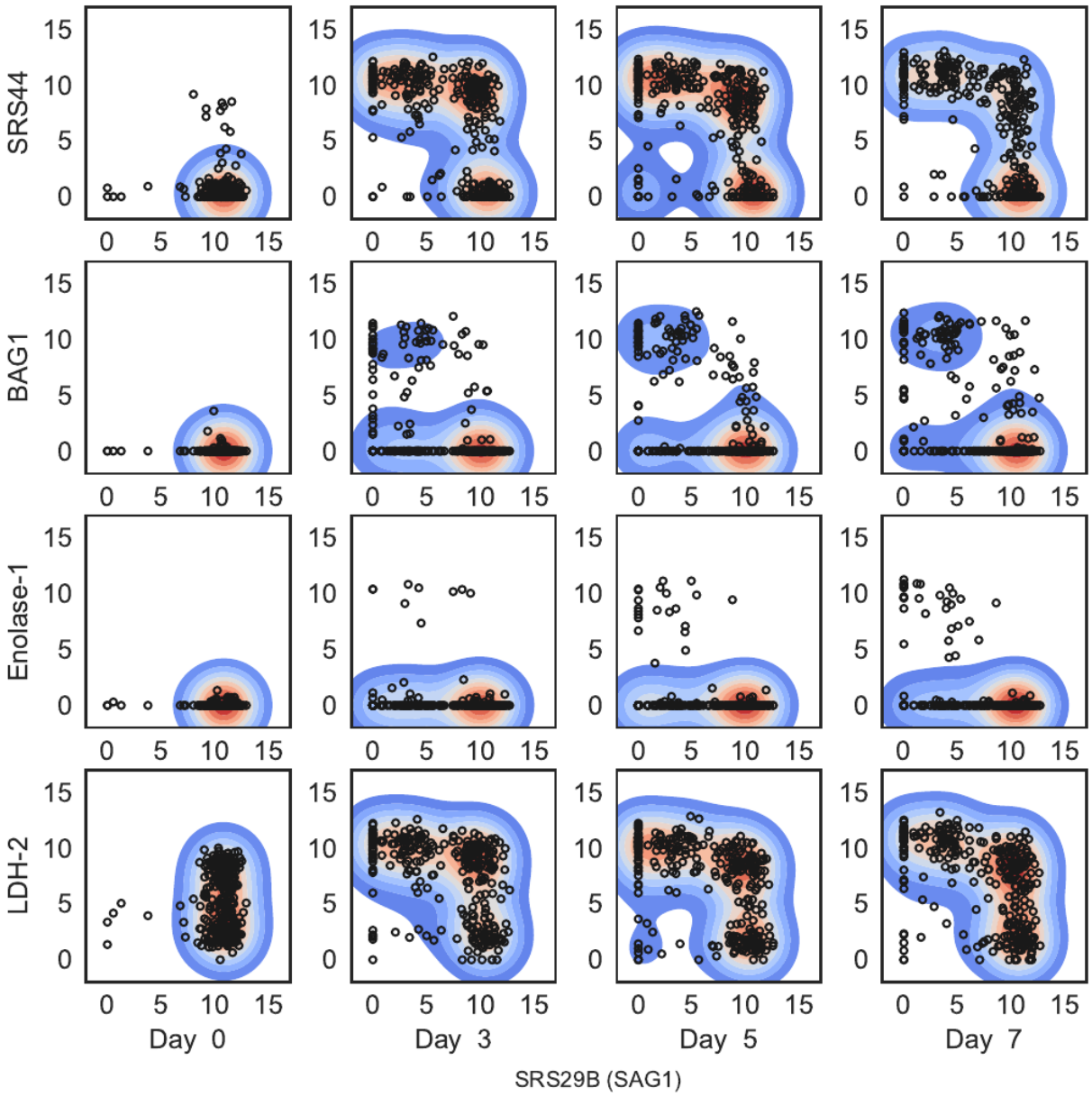

Figure 3 - Supplementary Figure 5. Expression level ($\log_2$ CPM) of four "bradyzoite-specific" marker genes compared to that of "tachyzoite-specific" marker gene, *SAG1*, stratified by days post induction (dpi; columns).

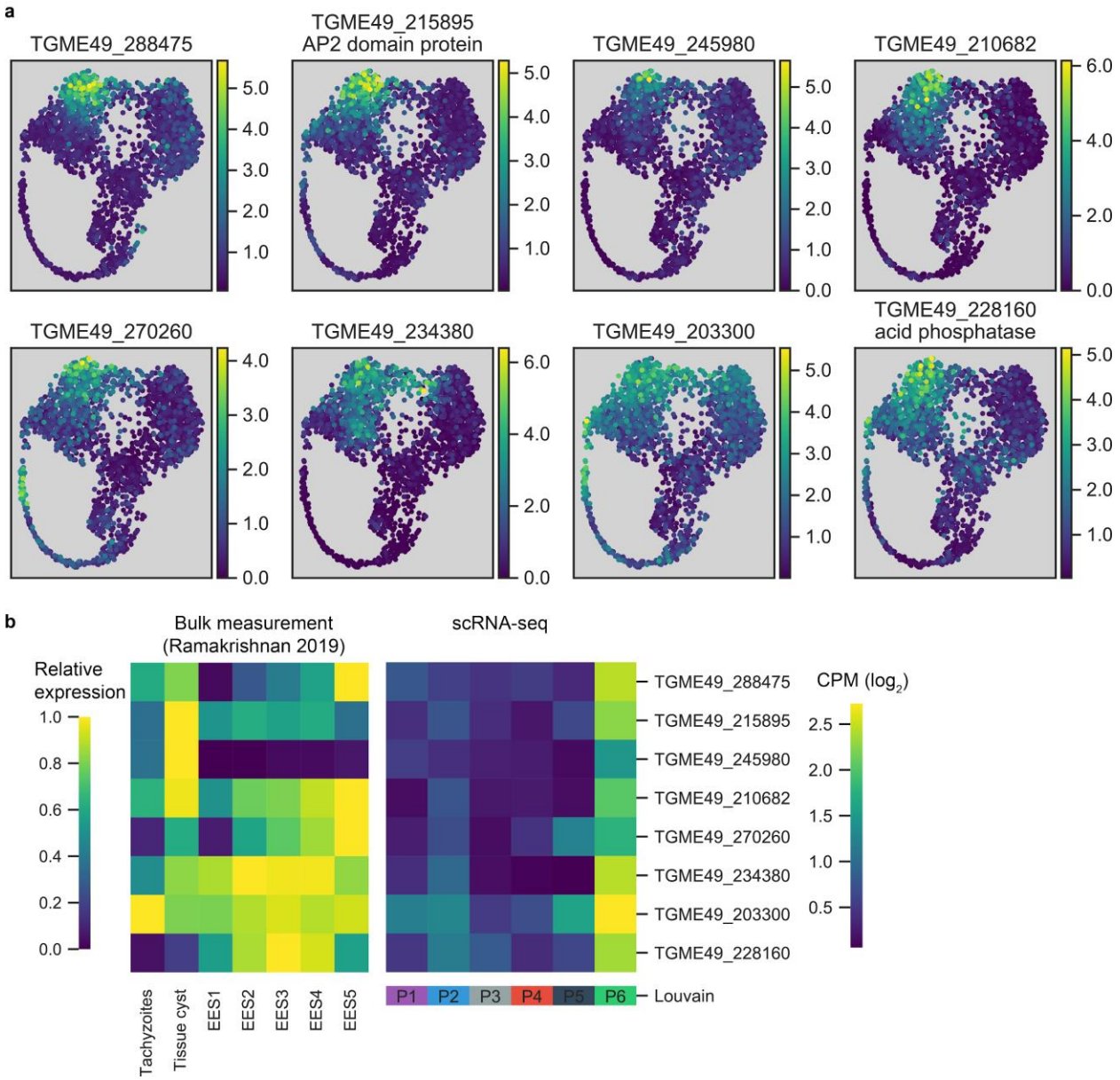

Figure 3 - Supplementary Figure 6. (a) UMAP projections of Pru colored by the neighbor-averaged expression (log_2 CPM) of top 8 most differentially expressed and enriched genes in P6 cluster relative to P1 and P2, two most closely related clusters. **(b)** Comparison of P6-specific genes in bulk measurement of tachyzoites, tissue cysts, or enteroepithelial stages (EES1-EES5)[43] (left) and scRNA-seq of Pru Louvain clusters (right).

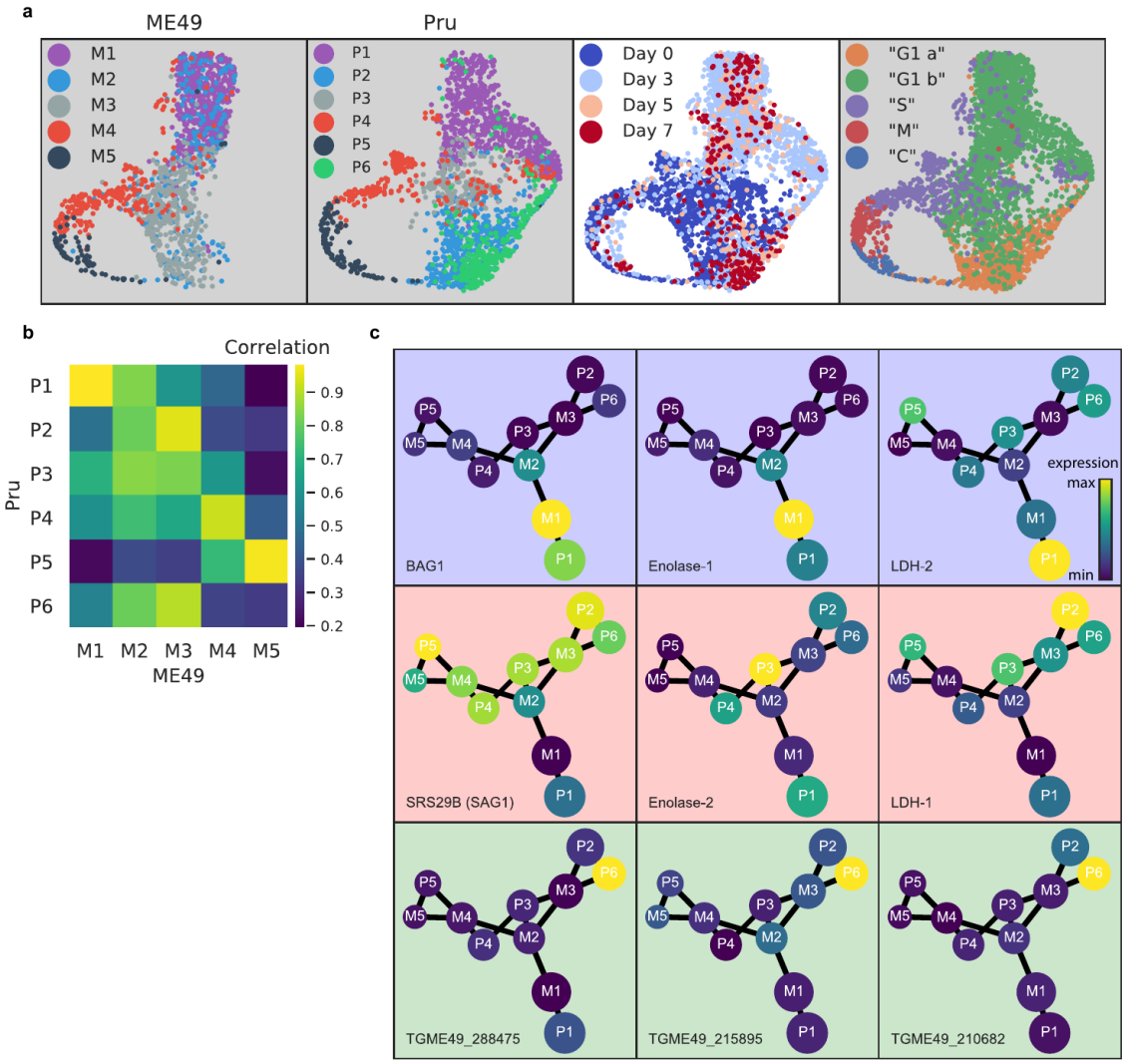

Figure 3 - Supplementary Figure 7. (a) UMAP projections of Pru and ME49 aligned by Scanorama. Cluster assignment was performed independently in each dataset. **(b)** Matrix correlation of cluster averaged expression between Pru and ME49. **(c)** Partition-based graph abstraction (PAGA) of aligned clusters with each being represented as a node connected by linkage with a connectivity threshold of 0.8. Node size reflects relative abundance of the cluster. Node colors reflect relative expression level ($\log_2$ CPM) of gene denoted in the bottom left of each panel, normalized to the maximum cluster expression of corresponding data set.

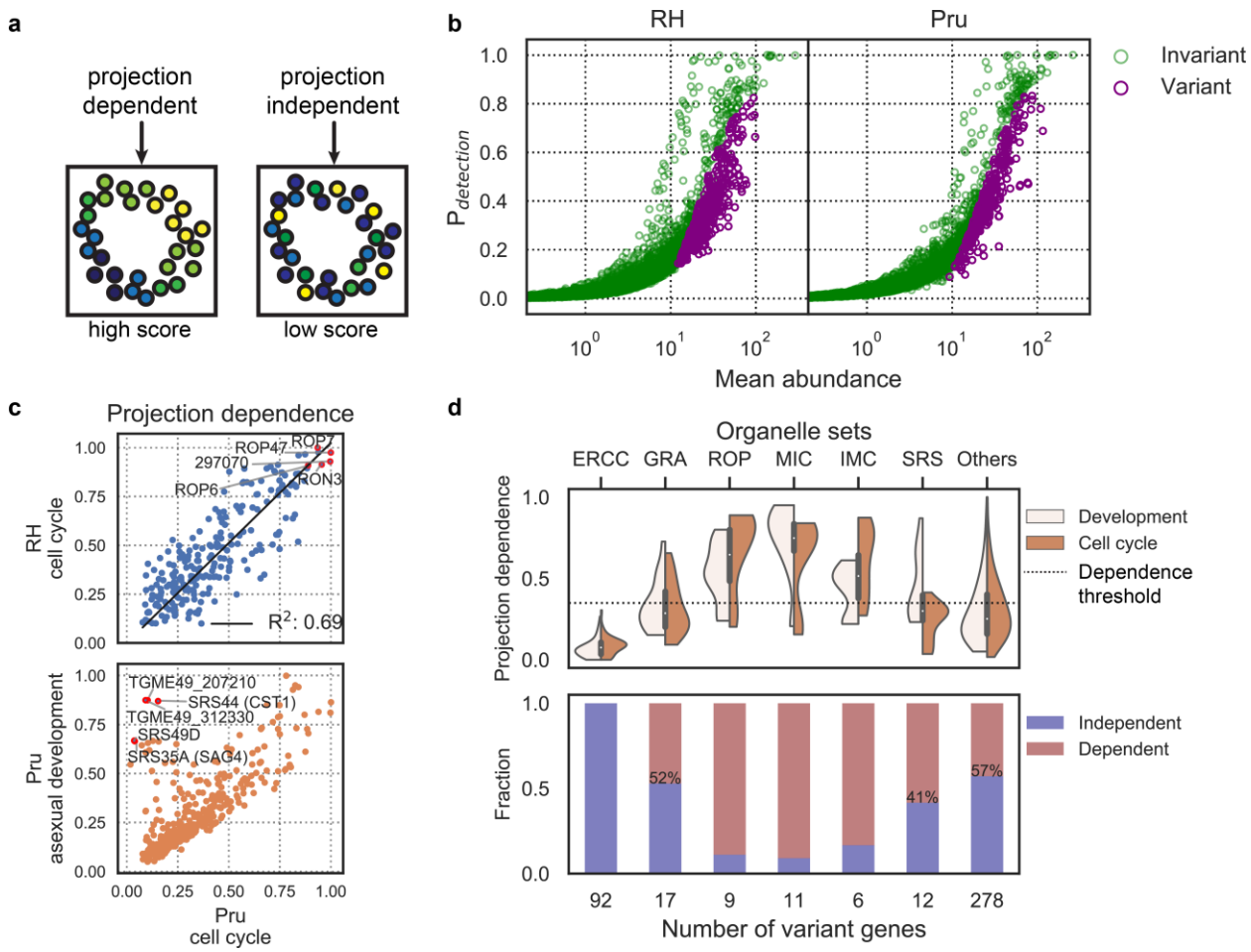

Figure 4. (a) Illustration of projection-dependent (left) and -independent (right) gene expression patterns. The spatial asymmetry is abolished after K-nearest neighbor (KNN) averaging of a projection-independent pattern, but not a dependent one. Projection-dependence score reflects normalized statistical significance of differences of observed KNN-averaged expression from a randomly permuted null distribution. Higher score indicates greater spatial asymmetry of the observed expression values. **(b)** Variant (purple) and invariant (green) genes are determined by identifying genes with detection rates lower than logistic regression model prediction evaluated by standard score test assuming one-sided Gaussian distribution ($p < 0.05$). **(c)** Top panel: comparison of RH and Pru (Day 0) cell cycle projection scores for intersecting variant genes in both RH and Pru. Linear regression fit (black solid line) is computed and the coefficient of determination (R^2) is reported on the top left corner. Examples of known ROP genes are shown. Bottom panel: Asexual development (UMAP) and cell cycle (PCA) projection scores for all variant genes in Pru (Day 0 - 7). Examples of genes with high dependence on development but not cell cycle are indicated. **(d)** Violin plots showing distribution of projection scores (top panel) and bar chart showing the fraction of dependent and independent genes (bottom panel) for all variant organelle sets in Pru. We identified 52%, 41%, and 57% independent genes amongst GRA, SRS, and Others (non-parasite-specific) gene sets, respectively.

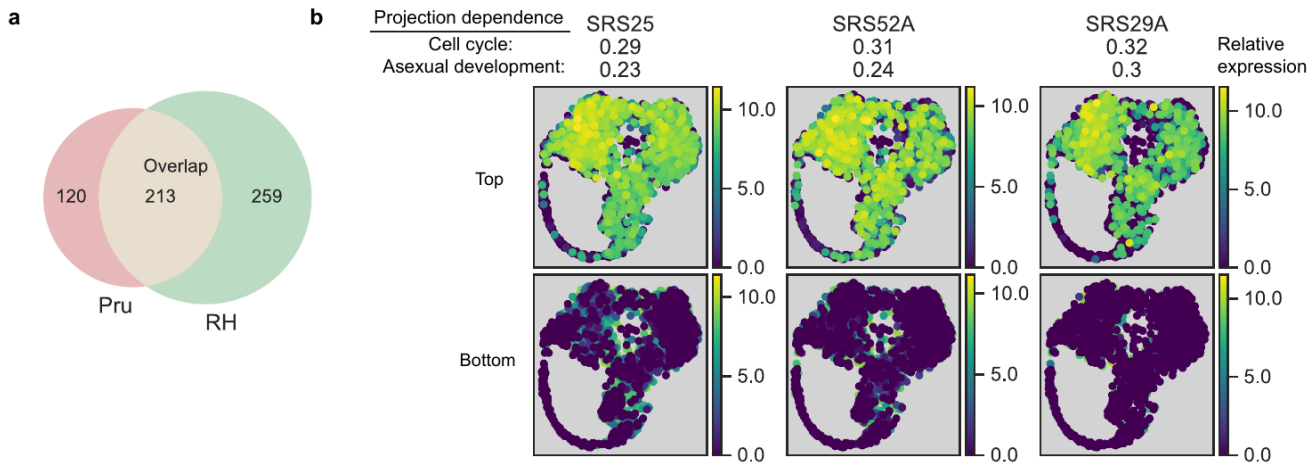

Figure 4 - Supplementary Figure 8. (a) Venn diagram showing the intersect and disjoint of variant genes identified in Pru (Day 0) and RH. **(b)** Expression level ($\log_2$ CPM) of SRS genes with low dependence for both cell cycle and asexual development, with projection scores reported above the panels. The cells are placed in ascending order of gene expression from top of the plot (upper panels) to highlight the cells with highest expression or from the bottom of the plot (lower panels) to show cells with the lowest expression. The two views reveal that cells with high and low expression can be neighbors and have weak correlation with the projection space; hence, these genes show very little dependence of gene expression on either cell cycle or asexual development which together drive the projection.

References

1. Pappas, G., Roussos, N. & Falagas, M. E. Toxoplasmosis snapshots: Global status of Toxoplasma gondii seroprevalence and implications for pregnancy and congenital toxoplasmosis. *Int. J. Parasitol.* (2009). doi:10.1016/j.ijpara.2009.04.003
2. Rabaud, C. *et al.* Extracerebral toxoplasmosis in patients infected with hiv a french national survey. *Med. (United States)* (1994). doi:10.1097/00005792-199411000-00004
3. Robert-Gangneux, F. *et al.* Molecular diagnosis of toxoplasmosis in immunocompromised patients: A 3-year multicenter retrospective study. *J. Clin. Microbiol.* (2015). doi:10.1128/JCM.03282-14
4. Sutterland, A. L. *et al.* Beyond the association. Toxoplasma gondii in schizophrenia, bipolar disorder, and addiction: Systematic review and meta-analysis. *Acta Psychiatr. Scand.* (2015). doi:10.1111/acps.12423
5. Vyas, A., Kim, S.-K., Giacomini, N., Boothroyd, J. C. & Sapolsky, R. M. Behavioral changes induced by Toxoplasma infection of rodents are highly specific to aversion of cat odors. *Proc. Natl. Acad. Sci.* (2007). doi:10.1073/pnas.0608310104
6. Soête, M., Camus, D. & Dubrametz, J. F. Experimental induction of bradyzoite-specific antigen expression and cyst formation by the RH strain of Toxoplasma gondii in vitro. *Exp. Parasitol.* (1994). doi:10.1006/expr.1994.1039
7. Jeffers, V., Tampaki, Z., Kim, K. & Sullivan, W. J. A latent ability to persist: differentiation in Toxoplasma gondii. *Cellular and Molecular Life Sciences* (2018). doi:10.1007/s00018-018-2808-x
8. Buchholz, K. R. *et al.* Identification of Tissue Cyst Wall Components by Transcriptome Analysis of In Vivo and In Vitro Toxoplasma gondii Bradyzoites . *Eukaryot. Cell* (2011). doi:10.1128/ec.05182-11
9. Manger, I. D. *et al.* Expressed Sequence Tag Analysis of the Bradyzoite Stage of Toxoplasma gondii: Identification of Developmentally Regulated Genes. *Infect. Immun.* (1998).
10. Pittman, K. J., Aliota, M. T. & Knoll, L. J. Dual transcriptional profiling of mice and Toxoplasma gondii during acute and chronic infection. *BMC Genomics* (2014). doi:10.1186/1471-2164-15-806
11. Yip, K. Our implementation of the SCA method. 1–4 (2007).
12. Cleary, M. D., Singh, U., Blader, I. J., Brewer, J. L. & Boothroyd, J. C. Toxoplasma gondii Asexual Development: Identification of Developmentally Regulated Genes and Distinct Patterns of Gene Expression . *Eukaryot. Cell* (2002). doi:10.1128/ec.1.3.329-340.2002
13. Radke, J. R. *et al.* The transcriptome of Toxoplasma gondii. *BMC Biology* (2005). doi:10.1186/1741-7007-3-26
14. Chen, L. F. *et al.* Comparative studies of Toxoplasma gondii transcriptomes: Insights into stage conversion based on gene expression profiling and alternative splicing. *Parasites and Vectors* (2018). doi:10.1186/s13071-018-2983-5
15. Fouts, A. E. & Boothroyd, J. C. Infection with Toxoplasma gondii bradyzoites has a diminished impact on host transcript levels relative to tachyzoite infection. *Infect. Immun.* (2007). doi:10.1128/IAI.01228-06
16. Hong, D.-P., Radke, J. B. & White, M. W. Opposing Transcriptional Mechanisms Regulate Toxoplasma Development . *mSphere* (2017). doi:10.1128/msphere.00347-16
17. White, M. W., Radke, J. R. & Radke, J. B. Toxoplasma development - turn the switch on or off? *Cell. Microbiol.* **16**, 466–472 (2014).
18. Soete, M., Fortier, B., Camus, D. & Dubremetz, J. F. Toxoplasma gondii: Kinetics of bradyzoite-tachyzoite interconversion in vitro. *Exp. Parasitol.* (1993). doi:10.1006/expr.1993.1031
19. Watts, E. *et al.* Novel Approaches Reveal that Toxoplasma gondii Bradyzoites within

Tissue Cysts Are Dynamic and Replicating Entities In Vivo. *MBio* (2015). doi:10.1128/mbio.01155-15

20. Ferguson, D. J. P., Huskinson-Mark, J., Araujo, F. G. & Remington, J. S. A morphological study of chronic cerebral toxoplasmosis in mice: comparison of four different strains of Toxoplasma gondii. *Parasitol. Res.* (1994). doi:10.1007/BF00932696
21. Radke, J. R., Guerini, M. N., Jerome, M. & White, M. W. A change in the premitotic period of the cell cycle is associated with bradyzoite differentiation in Toxoplasma gondii. *Mol. Biochem. Parasitol.* (2003).
22. Sinai, A. P. *et al.* Reexamining Chronic Toxoplasma gondii Infection: Surprising Activity for a “Dormant” Parasite. *Current Clinical Microbiology Reports* (2016). doi:10.1007/s40588-016-0045-3
23. Jerome, M. E., Radke, J. R., Bohne, W., Roos, D. S. & White, M. W. Toxoplasma gondii bradyzoites form spontaneously during sporozoite- initiated development. *Infect. Immun.* (1998).
24. Ali, F. *et al.* Cell cycle-regulated multi-site phosphorylation of Neurogenin 2 coordinates cell cycling with differentiation during neurogenesis. *Development* (2011). doi:10.1242/dev.067900
25. Kim, D. H. *et al.* The CRL4Cdt2 Ubiquitin Ligase Mediates the Proteolysis of Cyclin-Dependent Kinase Inhibitor Xic1 through a Direct Association with PCNA. *Mol. Cell. Biol.* (2010). doi:10.1128/mcb.01135-09
26. Radke, J. R. & White, M. W. A cell cycle model for the tachyzoite of Toxoplasma gondii using the Herpes simplex virus thymidine kinase. *Mol. Biochem. Parasitol.* (1998). doi:10.1016/S0166-6851(98)00074-7
27. Conde de Felipe, M. M., Lehmann, M. M., Jerome, M. E. & White, M. W. Inhibition of Toxoplasma gondii growth by pyrrolidine dithiocarbamate is cell cycle specific and leads to population synchronization. *Mol. Biochem. Parasitol.* (2008). doi:10.1016/j.molbiopara.2007.09.003
28. Behnke, M. S. *et al.* Coordinated progression through two subtranscriptomes underlies the tachyzoite cycle of toxoplasma gondii. *PLoS One* **5**, (2010).
29. Wang, B. *et al.* Stem cell heterogeneity drives the parasitic life cycle of Schistosoma mansoni. 1–23 (2018).
30. Reid, A. J. *et al.* Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. *Elife* **7**, 1–29 (2018).
31. Ngara, M. *et al.* Exploring parasite heterogeneity using single-cell RNA-seq reveals a gene signature among sexual stage Plasmodium falciparum parasites. *Exp. Cell Res.* (2018). doi:10.1016/j.yexcr.2018.08.003
32. Poran, A. *et al.* Single-cell RNA sequencing reveals a signature of sexual commitment in malaria parasites. *Nature* **551**, 95–99 (2017).
33. Svensson, V. *et al.* Power analysis of single-cell rnA-sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).
34. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell* **65**, 631-643.e4 (2017).
35. Reid, A. J. Large, rapidly evolving gene families are at the forefront of host-parasite interactions in Apicomplexa. *Parasitology* (2014). doi:10.1017/S0031182014001528
36. Lönnberg, T. *et al.* Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves $T_H 1/T_{FH}$ fate bifurcation in malaria. *Sci. Immunol.* **2**, eaal2192 (2017).
37. Manno, G. La *et al.* RNA velocity in single cells. *bioRxiv* 206052 (2017). doi:10.1101/206052
38. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 1–5 (2018).

39. Beraki, T. *et al.* Divergent kinase regulates membrane ultrastructure of the Toxoplasma parasitophorous vacuole . *Proc. Natl. Acad. Sci.* (2019). doi:10.1073/pnas.1816161116
40. Coffey, M. J. *et al.* Aspartyl Protease 5 Matures Dense Granule Proteins That Reside at the Host-Parasite Interface in Toxoplasma gondii . *MBio* (2018). doi:10.1128/mbio.01796-18
41. Jones, N. G., Wang, Q. & Sibley, L. D. Secreted protein kinases regulate cyst burden during chronic toxoplasmosis. *Cell. Microbiol.* (2017). doi:10.1111/cmi.12651
42. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* (2018). doi:10.21105/joss.00861
43. Ramakrishnan, C. *et al.* An experimental genetically attenuated live vaccine to prevent transmission of Toxoplasma gondii by cats. *Sci. Rep.* (2019). doi:10.1038/s41598-018-37671-8
44. Hie, B. L., Bryson, B. & Berger, B. Panoramic stitching of heterogeneous single-cell transcriptomic data. *bioRxiv* (2018). doi:10.1101/371179
45. Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* (2019). doi:10.1186/s13059-019-1663-x
46. Lyons, R. E., McLeod, R. & Roberts, C. W. Toxoplasma gondii tachyzoite-bradyzoite interconversion. *Trends in Parasitology* (2002). doi:10.1016/S1471-4922(02)02248-1
47. Camejo, A. *et al.* Identification of three novel Toxoplasma gondii rhoptry proteins. *Int. J. Parasitol.* (2014). doi:10.1016/j.ijpara.2013.08.002
48. Weiss, L. M., Ma, Y. F., Takvorian, P. M., Tanowitz, H. B. & Wittner, M. Bradyzoite development in Toxoplasma gondii and the hsp70 stress response. *Infect. Immun.* (1998).
49. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* (2015). doi:10.1038/nrg3833
50. Grün, D. & Van Oudenaarden, A. Design and Analysis of Single-Cell Sequencing Experiments. *Cell* (2015). doi:10.1016/j.cell.2015.10.039
51. Saeij, J. P. J. *et al.* Toxoplasma co-opts host gene expression by injection of a polymorphic kinase homologue. *Nature* **445**, 324–327 (2007).
52. Kim, S.-K. & Boothroyd, J. C. Stage-Specific Expression of Surface Antigens by Toxoplasma gondii as a Mechanism to Facilitate Parasite Persistence. *J. Immunol.* (2005). doi:10.4049/jimmunol.174.12.8038
53. WEISS, L. M. *et al.* A Cell Culture System for Study of the Development of Toxoplasma gondii Bradyzoites. *J. Eukaryot. Microbiol.* (1995). doi:10.1111/j.1550-7408.1995.tb01556.x
54. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* (2014). doi:10.1038/nprot.2014.006
55. Tarashansky, A. J., Xue, Y., Quake, S. R. & Wang, B. Self-assembling Manifolds in Single-cell RNA Sequencing Data. *bioRxiv* (2018). doi:10.1101