Running head: Single-cell RNA-Seq of *Tetrahymena*

Title: Revealing Gene Expression Heterogeneity in a Clonal Population of *Tetrahymena thermophila* through Single-Cell RNA Sequencing

Authors: Hiroki Kojima[a], Akiko Kashiwagi[b,c], Takashi Ikegami[a]

a. Graduate School of Arts and Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan

b. Faculty of Agriculture and Life Science, Hirosaki University, Bunkyo-cho 3, Hirosaki, Aomori, 030-8561, Japan

c. The United Graduate School of Agricultural Science, Iwate University, Ueda-3, Morioka, Iwate, 020-8550, Japan

Correspondence

H. Kojima, Graduate School of Arts and Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan

e-mail:kojima@sacral.c.u-tokyo.ac.jp

Abstract
We investigated the gene expression variability in a population of 5,000 *Tetrahymena thermophila* cells using single-cell RNA sequencing. During preprocessing, we enhanced read mappings by estimating the 3′-untranslated regions using the "peaks2utr" method. The resulting expression patterns enabled us to identify several prevalent clusters within the RNA expression space. Some of these clusters corresponded to specific cell phases. After regressing out these phase-related clusters, we performed a gene ontology analysis to identify the remaining clusters. We propose that these remaining clusters represent emerging sub-populations within the *Tetrahymena* population, primarily associated with translation, energy production, and cell structure.

Keywords: ciliates, 3′-untranslated region estimation, cell-cycle dependent genes, gene expression variability, gene ontology analysis

Introduction

Gene expression variability is a widely observed phenomenon in both prokaryotes and eukaryotes (Raj and van Oudenaarden, 2008; Raser and O'Shea, 2005). Single-cell RNA sequencing (scRNA-seq) provides a powerful tool to dissect the mechanisms behind this variability (Saliba et al., 2014; Buettner et al., 2015; Kuchina et al, 2021). While mammalian cells have been extensively analyzed using scRNA-seq, analogous studies on ciliates, particularly those involving sequencing thousands of cells, have been comparatively rare, despite previous bulk RNA-seq and transcriptome analysis of ciliates (Xiong et al., 2012).

One reason for this disparity is the larger size of ciliate cells (~500 μm in length) compared to mammalian cells. Most methods that enable scRNA-seq of thousands of cells are droplet-based, which imposes limitations on the size of cells that can be assessed (~40 μm) (Ziegenhain et al., 2017).

In this study, we present the results of a scRNA-seq analysis of a ciliate, *Tetrahymena thermophila*, conducted to investigate gene expression variability. We chose *T. thermophila* as our subject not only because it serves as a primary ciliate model organism for molecular biology and genetics (Turkewitz et al., 2002), but also due to its several advantageous characteristics. Firstly, *T. thermophila* is relatively small, with dimensions of approximately 40–60 μm in length and 20–30 μm in width, making it suitable for droplet-based scRNA-seq techniques. Secondly, *T. thermophila* mRNA is transcribed from the macronucleus (MAC), and the complete genome sequences of the MAC have already been reported (Eisen et al., 2006; Wang et al, 2021). Lastly, *T. thermophila* has been observed to exhibit individual variability in behavior (Jordan et al., 2013).

For single-cell RNA sequencing (scRNA-seq), we adopted the 10x Genomics 3' gene expression analysis method. This choice was guided by the fact that approximately 90% of the reads from total RNA from a single *T. thermophila* cell were reported to be from the large subunit rRNA (LSU rRNA) (Kolisko et al., 2014). The 10x Genomics 3' gene expression analysis is a polyadenylated (poly-A) tail capture method, which effectively minimizes contamination from reads originating from LSU rRNA.

However, this method has some limitations, particularly in the context of *T. thermophila*. The genome sequence and annotation information for *T. thermophila* are relatively less accurate compared to those for mammalian cells. This issue is crucial, especially considering the annotation of the 3' untranslated region (3' UTR). This is due to the fact that approximately 65% of the reads obtained using the 10x Single Cell 3' v3 chemistry were reported to map to both 3' UTRs and exon regions (Zhang et al., 2019), whereas in bulk RNA-seq many reads map on the structural genes.

To tackle this issue, we utilized the sequence data we obtained to refine the annotation of the 3' UTR using the peaks2utr method (Haese-Hill et al., 2022). This approach allowed us to augment the precision of the 3' UTR annotation, which subsequently enhanced the results of the read mappings.

We then leveraged these scRNA-seq results to probe the heterogeneity of the *T. thermophila* population. Variability in gene expression can be indicative of both stochasticity in gene expression and heterogeneity in individual cell states. Although previous studies suggest that the majority of gene expression variability stems from differences in cell states such as cell volume, cell cycle, differentiation, and calcium response (Foreman and Wollman,

2020), it has been proposed that variations in cell cycle phase could be a primary contributing factor. Lacking well-established markers for each phase of the cell cycle in *T. thermophila*, we utilized a recently reported set of 3,244 genes with cell-cycle-dependent expression patterns (Zhang et al., 2023). We selected 996 of these genes to infer the cell cycle phase of each individual cell, allowing us to regress out the variations attributable to cell cycle phases and reveal heterogeneity from other sources.

In summary, we employed the 10x Genomics system to conduct scRNA-seq on approximately 5,000 *T. thermophila* cells. To discern the effect of cell-cycle-dependent genes, we performed regression analysis to remove their influence from the gene expression data. Upon visualizing the cells using uniform manifold approximation and projection (UMAP), several distinct clusters became apparent. These clusters represented diverse subpopulations within the *T. thermophila* population, shedding light on the underlying transcriptional variability at the single-cell level.

MATERIALS AND METHODS

The detailed experimental procedures and analysis methods are described in the supplementary text.

Preparation of libraries for scRNA-seq

The *T. thermophila* SB210-E strain was used from the Tetrahymena Stock Center (Cornell University, USA). Cells in the exponential growth phase were loaded onto a 10x Chromium platform using the 10x Single Cell 3′ v3 chemistry, targeting 5,000 cells. Libraries were sequenced with NextSeq2000 (Illumina, San Diego, CA).

Analysis of the sequencing data

We processed the sequencing data using the 10x Cell Ranger pipeline, version 7.0.1. The reference annotations of the MAC sequence were JCVI-TTA1-2.2 (National Center for Biotechnology Information [NCBI] USA) and T. thermophila _MAC_2021-Updated (Tetrahymena Genome Database [TGD]) (Stover et al., 2006), and we refined the 3′-UTR annotation using the peaks2utr method (Haese-Hill et al., 2022).

The expression matrix obtained from Cell Ranger was processed using Seurat (Hao et al., 2021) and normalized by sctransform (Hafemeister and Satija, 2019). For regressing out the cell phase effect, we used the markers of each cell phase (Table S2) selected based on the list from Zhang et al. 2023.

Data availability

The FASTQ file and analyzed file can be found in DDBJ with the accession numbers DRR440460 and E-GEAD-611, respectively.

RESULTS AND DISCUSSION

Mapping of reads

Upon annotating the 3' UTR for the two MAC sequences available in the NCBI database and the TGD, we mapped the reads to them. We newly annotated 16,265 and 16,164 3' UTRs for the *T. thermophila* genome sequences in the NCBI and TGD databases, respectively. The percentage of reads confidently mapped to the genome was 81.3% and 14.4% for NCBI and TGD sequences, respectively (Table S1). The low percentage for the TGD sequence indicated that it did not cover the entire genome; hence, we used the NCBI database sequence for further analysis. While the genome mapping was high at 81.3%, the transcriptome mapping for the NCBI database was only 9.5%. The percentage of reads confidently mapped to intergenic regions in our result was 71.0%, indicating that there are still many unannotated regions in the NCBI database.

Clustering results

The expression patterns were divided into nine clusters by analyzing the outputs from Cell Ranger using Seurat (Fig.1(a) and (c)). Each cluster was characterized by a set of genes exhibiting unique high expression within that cluster (Fig.1(c)). Each cluster contained between 59 to 202 significantly upregulated genes, though in clusters 0 and 5, only two (elongation factor Tu and hypothetical protein) and four genes (immobilization antigen and papain family cysteine protease) were significantly upregulated, respectively.

To characterize each cluster, GO enrichment analysis was performed for seven clusters excluding clusters 0 and 5 (Figure S1 (a)). Cluster 1 exhibited high expression of genes related to translation and cluster 2 displayed elevated expression of genes associated with translation, as well as the carboxylic acid metabolic process. Cluster 6 showed enrichment in processes related to microtubule-based activities and microtubule movement including dyneins. Upon analyzing the top 30 significantly upregulated genes within cluster 6, we identified genes associated with cell division, such as *atu1*, *nudF*, and the nuclear movement protein gene. These findings suggest that cluster 6 probably contained cells in the MAC-amitosis phase.

When analyzing the top 15 significantly upregulated genes within cluster 3, we found that nine of them were histone genes, including *htb2*, *hht2*, *hho1*, *hta2*, *htb1*, *hhf1*, *hhf2m*, and *hta1*, and the *histone H3.4* gene. Additionally, two genes associated with DNA bending or supercoiling, *hmb2*, and *hmbc*, were identified. Notably, *hhf1*, *hhf2*, *hmb1*, and *hmb2* are registered as genes expressed during the MAC-S phase in TGD. These findings suggested that cluster 3 probably consisted of cells in the MAC-S phase.

Next, we analyzed the top 30 significantly upregulated genes within cluster 4, 7, and 8. Both clusters 4 and 7 contained genes of tetrins located in the oral apparatus, but with distinct types of calcium-binding proteins: EF-hand pair proteins for cluster 4 and IQ calmodulin-binding motif proteins for cluster 7. In cluster 8, we identified putative transmembrane proteins and calmodulin-related proteins.

Genes that had been previously reported as cell cycle-dependent (Zhang et al., 2023) were mapped on UMAP (Figure S2 (a)). Cells expressing genes for MAC-late S phase were

found in cluster 3, whereas cells expressing genes for MAC-amitosis phase were found in cluster 6. Cells expressing MAC-G1 and G2 phase genes were distributed across various clusters on UMAP (Figure S2 (a)), indicating that the expression pattern of MAC-G1 and -G2 phase genes differed among individual cells.

Regressing out the cell cycle scores

We observed subpopulations after removing the effects of the cell cycle using the function provided in Seurat (Hao et al., 2021). Cell populations analyzed via scRNA-seq comprise cells in various cell-cycle phases. Therefore, the cell cycle-dependent gene expression can significantly bias the analysis and clustering of our data. After regressing out the cell cycle effect, the expression pattern was divided into eight clusters (Fig. 1(b) and (d)). Mapping of cells that expressed the MAC-late S and MAC-amitosis phase genes on UMAP showed that these were scattered across different clusters, indicating that the regression had been successful (Fig. S2 (b)). These results clearly showed that diverse subpopulations were present within the population.

GO enrichment analysis was conducted to characterize each cluster, except for clusters 0 and 4, as these only consisted of one (elongation factor Tu) and four genes (immobilization antigen and papain family cysteine protease), respectively (Fig. S1(b)). Clusters 1, 2, 5, and 6 included genes relating to translation and/or carboxylic acid metabolic process (Fig. S2 (b)). However, numerous genes significantly upregulated in clusters 3, 5, 6, and 7 were either unclassified by GO or designated as hypothetical proteins. Upon analyzing the top 30 significantly upregulated genes within these clusters, we identified genes associated with oral organs, including tetrins in cluster 3, calcium binding proteins, including calmodulin and EF-hand protein, in cluster 5, polyubiquitin that is post-translation modification protein in cluster 6, and dynein proteins that are cilia formation in cluster 7. These results indicated the presence of functionally different subpopulations in the population.

In our study, several clusters remained after regressing out the effect of cell phases. Traditionally, the internal state of a cell is considered to be uniform during the exponential growth phase of microorganisms that undergo binary fission as the specific growth rate remains constant (Madigan et al., 2021). However, recent scRNA-seq studies on *Bacillus subtilis* have provided clear evidence that, even during the exponential phase, gene expression patterns in *B. subtilis* undergo changes as the cells grow (Kuchina et al., 2021). The application of scRNA-seq in the study of microorganisms is crucial as it introduces new insights to the field of eukaryotic microbiology, and underscores the value of scRNA-seq as a tool for revealing the complex and dynamic nature of cellular populations.

ACKNOWLEDGEMENTS

This study was supported in part by JSPS KAKENHI Grant Number 21H04885 and UEHARA foundation. The authors thank Dr. Geoffrey Kapler for providing a list of cell cycle-dependent genes.

FIGURE LEGEND

Figure 1. scRNA-seq clustering analysis

(a, b) UMAP representation of the RNA expression pattern before and after regression of cell cycle-dependent genes, respectively. (c, d) Heat maps showing differential expression in the population before and after regression of cell cycle dependent genes, respectively.

REFERENCES

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C. & Stegle, O. 2015. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol*., 33: 155-160.

Eisen, J. A., Coyne, R. S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J. R., Badger, J. H., Ren, Q., Amedeo, P., Jones, K. M., Tallon, L. J., Delcher, A. L., Salzberg, S. L., Silva, J. C., Haas, B. J., Majoros, W. H., Farzad, M., Carlton, J. M., Smith, R. K., Jr., Garg, J., Pearlman, R. E., Karrer, K. M., Sun, L., Manning, G., Elde, N. C., Turkewitz, A. P., Asai, D. J., Wilkes, D. E., Wang, Y., Cai, H., Collins, K., Stewart, B. A., Lee, S. R., Wilamowska, K., Weinberg, Z., Ruzzo, W. L., Wloga, D., Gaertig, J., Frankel, J., Tsao, C. C., Gorovsky, M. A., Keeling, P. J., Waller, R. F., Patron, N. J., Cherry, J. M., Stover, N. A., Krieger, C. J., del Toro, C., Ryder, H. F., Williamson, S. C., Barbeau, R. A., Hamilton, E. P.& Orias, E. 2006. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS. Biol*., 4, e286. doi: 10.1371/journal.pbio.0040286.

Foreman, R. & Wollman, R. 2020. Mammalian gene expression variability is explained by underlying cell state. *Mol. Syst. Biol*., 16, e9146. doi: 10.15252/msb.20199146.

Haese-Hill, W., Crouch, K. & Otto, T. D. 2023. peaks2utr: a robust Python tool for the annotation of 3' UTRs. *Bioinformatics*, 39, btad112. doi:10.1093/bioinformatics/btad112.

Hafemeister, C. & Satija, R. 2019. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.*, 20, 296. doi:10.1186/s13059-019-1874-1

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., 3rd, Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., Rogers, A. J., McElrath, J. M., Blish, C. A., Gottardo, R., Smibert, P. & Satija, R. 2021. Integrated analysis of multimodal single-cell data. *Cell*, 184: 3573–3587.

Jordan, D., Kuehn, S., Katifori, E. & Leibler, S. 2013. Behavioral diversity in microbes and low-dimensional phenotypic spaces. *Proc. Natl. Acad. Sci. USA*, 110:14018-14023.

Kolisko, M., Boscaro, V., Burki, F., Lynn, D. H. & Keeling, P. J. 2014. Single-cell transcriptomics for microbial eukaryotes. *Curr. Biol*., 24:R1081-1082.

Kuchina, A., Brettner, L. M., Paleologu, L., Roco, C. M., Rosenberg, A. B., Carignano, A., Kibler, R., Hirano, M., DePaolo, R. W. & Seelig, G. 2021. Microbial single-cell RNA sequencing by split-pool barcoding. *Science*, 371:eaba5257. doi: 10.1126/science.aba5257.

Madigan, M. T., Aiyer, J., Buckley, D. H., Sattley, W. M. & Stahl, D. A. 2021. Microbial growth and its control. In: Madigan, M. T., Aiyer, J., Buckley, D. H., Sattley, W. M. &

Stahl, D. A. (ed.), Brock Biology of Microorganisms. 16th ed. Pearson Education, London. p.155

Raj, A. & van Oudenaarden, A. 2008. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135:216-226.

Raser, J. M. & O'Shea, E. K. 2005. Noise in gene expression: origins, consequences, and control. *Science*, 309:2010-2013.

Saliba, A. E., Westermann, A. J., Gorski, S. A. & Vogel, J. 2014. Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res.*, 42:8845-8860.

Stover, N. A., Krieger, C. J., Binkley, G., Dong, Q., Fisk, D. G., Nash, R., Sethuraman, A., Weng, S. & Cherry, J. M. 2006. *Tetrahymena* Genome Database (TGD): a new genomic resource for *Tetrahymena thermophila* research. *Nucleic Acids Res*., 34:D500-D503.

Turkewitz, A. P., Orias, E. & Kapler, G. 2002. Functional genomics: the coming of age for *Tetrahymena thermophila*. *Trends Genet*., 18:35-40.

Wang, X., He, Y., Zhang, Q., Ren, X. & Zhang, Z. 2021. Direct comparative analyses of 10X Genomics Chromium and Smart-seq2. *Genomics Proteomics Bioinformatics*, 19:253-266.

Xiong, J., Lu, X., Zhou, Z., Chang, Y., Yuan, D., Tian, M., Zhou, Z., Wang, L., Fu, C., Orias, E. & Miao, W. 2012. Transcriptome analysis of the model protozoan, *Tetrahymena thermophila*, using Deep RNA sequencing. *PLoS One*, 7, e30630. doi:10.1371/journal.pone.0030630.

Zhang, L., Cervantes, M. D., Pan, S., Lindsley, J., Dabney, A. & Kapler, G. M. 2023. Transcriptome analysis of the binucleate ciliate *Tetrahymena thermophila* with asynchronous nuclear cell cycles. *Mol. Biol. Cell*, 34, rs1, doi:10.1091/mbc.E22-08-0326.

Zhang, X., Li, T., Liu, F., Chen, Y., Yao, J., Li, Z., Huang, Y. & Wang, J. 2019. Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems. *Mol. Cell*, 73:30-142 e5, doi: 10.1016/j.molcel.2018.10.020.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I. & Enard, W. 2017. Comparative analysis of single-cell RNA sequencing methods. *Mol. Cell*, 65:631-643 e4. doi:10.1016/j.molcel.2017.01.023.

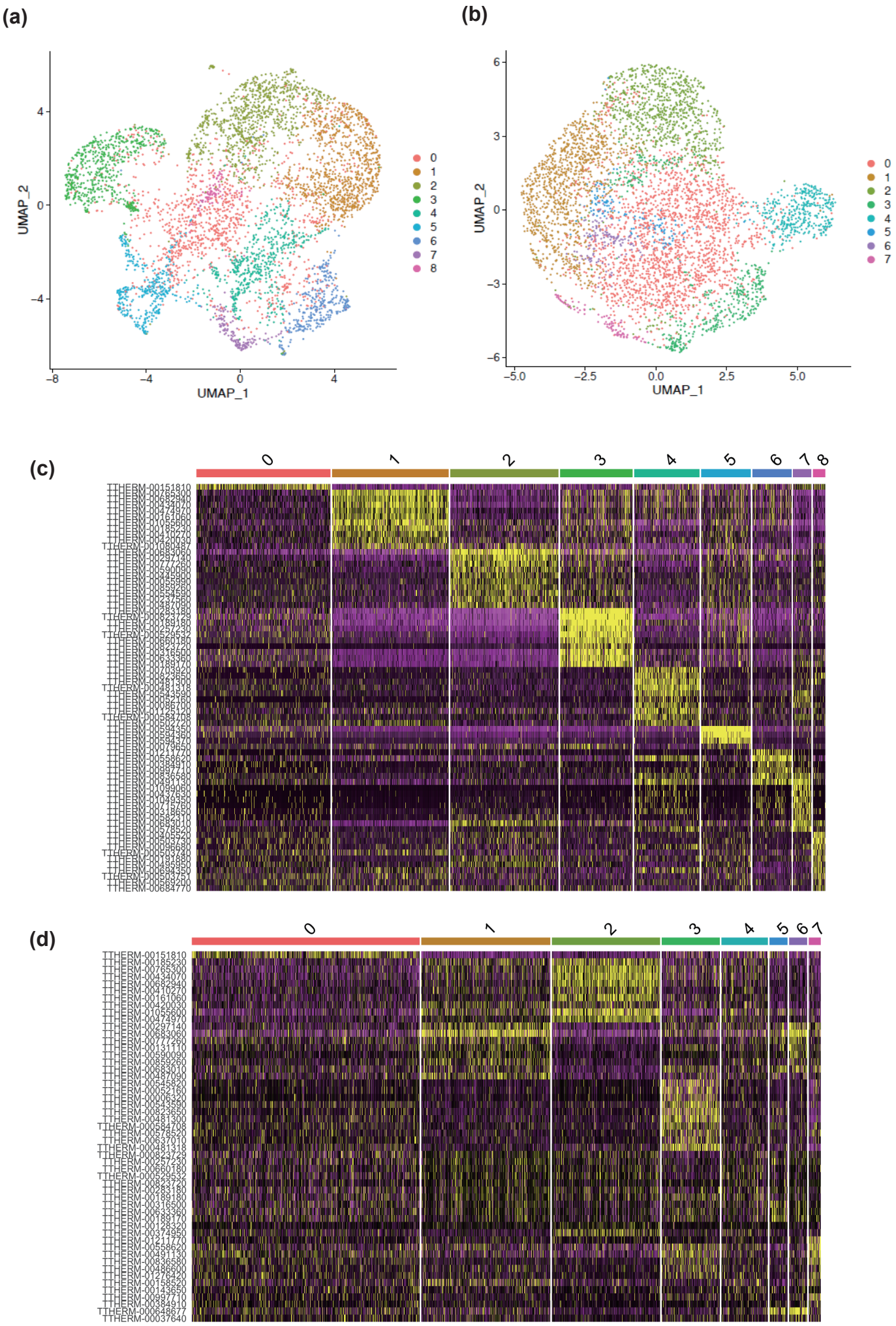
(a)
UMAP_2
UMAP_1
0
1
2
3
4
5
6
7
8
(b)
UMAP_2
UMAP_1
0
1
2
3
4
5
6
7
(c)
0
1
2
3
4
5
6
7
8
(d)
0
1
2
3
4
5
6
7

Figure 1