Investigating *Escherichia Coli* scRNA-seq data using Batch Integration Methods

BMEN 4480, Columbia University
Skylar Li, Jacob Rosenfeld

Abstract

Bacterial populations have long been known to show heterogeneity in gene expression for fitness and survival even in controlled environments. While bet hedging and division of labor are most proposed as the potential modulating factors, no consensus has been reached with respect to the mechanisms giving rise to variability in response to stress factors among single-cell organisms. As recent developments in scRNA-seq protocols in bacteria have overcome barriers such as low mRNA abundance and presence of cell walls, providing a small number of readily available scRNA-seq data, we evaluated the capabilities of established batch correcting algorithms available in the Python language environment to integrate *E. coli* batches exposed to different environmental conditions so to see more complex responses in differentiations in gene regulation. Imputation and batch-correction using single-cell variance inference (scVI) resulted in a good mixture of cells between both batches while preserving the underlying biological structure of the heat shock conditions and variable growth stages. Further downstream analysis indicated that cells exposed to heat shock are likely to employ different survival strategies through regulation of the methionine biosynthesis pathway, aligning with previous literature suggesting that *E. coli* utilize a survival tactic by diversifying survival mechanisms.

Introduction

Bacterial populations have long been known to show heterogeneity in gene expression for fitness and survival even in controlled environments. For instance, quorum sensing, the mechanism of regulating gene expression in response to cell-population density fluctuations through the production, release and detection of autoinducers, had previously been thought of as a collective behavior in which all members of a bacterial population partake. However, there has been an increase in reports of phenotypic heterogeneity both in autoinducer production and in target gene activation beyond variability due to microenvironmental differences. While bet hedging and division of labor are most proposed as the potential modulating factors, no consensus has been reached with respect to the molecular mechanisms nor prevalence of quorum sensing-related cell-to-cell variability [1]. Revolutionizing the field of transcriptional heterogeneity in diverse systems, single-cell RNA sequencing (scRNA-seq) assay methods for eukaryotic cells have become more readily available along with downstream analytic tools, with over 1000 tools cataloged in the scRNA-tools database. However, the application of scRNA-seq to prokaryotes has been delayed by their extremely low mRNA abundance and average copy number, lack of mRNA polyadenylation, interference of lysis due to cell walls, and small size hindering microfluidic single-cell isolation [2]. Although recent developments in scRNA-seq protocols in bacteria have been successfully benchmarked, there have been four main contributions to date, with two studies identifying *E. coli* transcripts [3][4] and two others finding Pseudomonas aeruginosa transcripts [5][6]. As access to bacterial single-cell transcriptomics is limited considering the considerable cost incurred by bacterial scRNA-seq protocols, the integration of datasets produced in different laboratories under different handling protocols could allow for further exploration of cell trajectories and responses. Additionally, integration can contribute to a more precise characterization of cellular states that are less discernible in noisy or incomplete data from one source alone [7]. Various batch correction methods can correct for systematic variations such as technical differences or experimental artifacts to achieve effective batch integration.

In this paper, two recently published *E. coli* scRNA-seq datasets are investigated using two different sequencing methods in order to elucidate gene expression differentiation at single cell resolution. In Blattman et. al (2020), a new method of scRNA-seq called Prokaryotic Expression-profiling by Tagging scRNA In Situ (PETRI-seq) using combinatorial indexing to barcode transcripts from tens of thousands of cells in a single experiment. In one of the experiments in Blattman et. al (2020), Gram Negative *E. coli* strain MG1655 cells are sequenced under variable conditions of growth including stationary and exponential growth stages by sequencing bacterial datasets after differing lengths of incubation. *E. coli* cells in the stationary growth phase are identified by the presence of red fluorescent protein (RFP) constructs while cells in the exponential growth phase are identified by green fluorescent protein (GFP) constructs. In Kuchina et. al (2021), a new method of scRNA-seq called Microbial Split-Pool Ligation Transcriptomics (microSPLiT) is designed which builds upon a eukaryotic scRNA-seq approach called SPLiT-seq labeling the cellular origin of RNA through combinatorial barcoding. In an experiment in Kuchina et. al (2021), Gram Positive B. subtilis PY79 and Gram Negative *E. coli* MW1255 species are mixed and exposed to heat shock and non-heat shock conditions [3][4].

The *E. coli* datasets are preprocessed and examined to confirm and replicate the findings of the original sources of the novel scRNA-seq prokaryotic datasets using 2D embeddings such as UMAP, tSNE, and principal component analysis (PCA) following library preparation. Cell clustering methods such as Leiden and Louvain clustering algorithms are applied to the individual, non-batch corrected datasets in order to investigate differentially expressed genes in the original data. Batch correction methods including Mutual Nearest Neighbors (MNN), Scanorama, and single cell variational inference (scVI) are applied to integrate the Blattman and Kuchina *E. coli* batches in order to be able to investigate and characterize differentially expressed genes and developmental trajectories of heterogeneous bacterial datasets as novel prokaryotic scRNA-seq methods continue to be developed in the foreseeable future. Our findings demonstrate the applicability of the current scRNA-seq analysis toolkit that has generally only had the opportunity to be applied in eukaryotic genomic analysis to date. By expanding our currently limited understanding of bacterial scRNA-seq, we hope to contribute to a granular understanding of cell responses, potentially contributing to breakthroughs in elucidating biofilm functionality, antibiotic resistance, persister cells, and other complex systems yet to be informed by bulk metagenomics.

Methods for batch correction

For Harmony, MNN, and ComBat, Scanorama, and scVI we conducted the data preprocessing steps of normalization, log-transformation, and scaling as documented by Kuchina et al. using the Scanpy package. For Scanorama and scVI, we further conducted data augmentation by identifying 921 genes present in both datasets of interest to use as a reference for gene-wise integration. To evaluate the batch correction results, we employed t-Distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP).

Mutual Nearest Neighbors (MNN)

MNN relies on three basic assumptions: at least one cell population is present in both batches, batch effects are orthogonal to the biological space, and the batch effect variations are smaller than biological effect variation between different cell types. Orthogonal batches are matched by mutual nearest neighbors (MNN) pairings by cell type in each batch subset. MNN pairings are identified by first conducting a global scaling through cosine normalization and calculating the Euclidean distance between pairs of cells. The MNN pairings are identified by finding k cells in an alternative batch nearest to a cell in the original batch and vice versa such that if two cells are mutually connected then they are an MNN pair. Finally, a cell-specific batch-correction vector weighted average of MNN pairs is computed with a Gaussian kernel for smoothness and collapses the alternative batch onto the original dataset [8].

Scanorama

Scanorama uses a panorama-like stitching technique to generalize mutual nearest neighbors and compress the gene expression profiles of each cell into a low-dimensional embedding using randomized SVD. Scanorama automatically identifies scRNA-seq datasets containing similar cells and can leverage only those matches for batch correction and integration. The algorithm uses an l2-normalization for expression values for each cell for scale invariant comparison and computed SVDs. Mutual nearest neighbors between datasets are computed and a non-linear batch-correction vector is determined in order to correct batches into a single integrated dataset. The output of Scanorama is a low-dimensional SVD representation of the original datasets that can be used for further downstream analysis [9].

scVI

Single cell variational inference (scVI) is a fully probabilistic approach that is based on a hierarchical Bayesian model with conditional distributions which can be trained using deep neural networks. The transcriptome of each cell is encoded into a low-dimensional latent vector of normal random variables which is then decoded to generate a posterior estimate of the original datasets. The transformation assumes a zero-inflated negative binomial distribution and can simultaneously perform imputation on the original datasets to predict missing genes and expression. The model can be trained on multiple datasets in order to remove batch effects and impute missing genes when creating a posterior estimate accounting for multiple datasets. Thus, scVI provides a computationally efficient method using a low-dimensional probabilistic representation in order to batch correct and impute scRNA-seq data [10].

Implementation and Results

Exponential and stationary growth stage distinction in Blattman et. al (2020)

In Blattman et. al (2020), the gram negative *E. coli* strain MW1655 are incubated at varying time periods in order to yield cells in the stationary and exponential growth phases. Plasmid constructs for green fluorescent protein (GFP) and red fluorescent protein (RFP) are used in order to identify exponential and stationary cells. Exponential growth *E. coli* cells are identified by the presence of GFP while stationary growth *E. coli* cells are identified by the presence of RFP. PETRI-seq was performed following incubation and mixing in order to develop a heterogeneous population of *E. coli* cells with naturally varying transcriptional profiles.

Original results were confirmed by filtering cells with gene expression lower than 15 counts and above 2000, normalized to median library size, log-transformed, and scaled the count matrix. Cells containing GFP, RFP, or none (did not capture expression of GFP or RFP) were identified and annotated for identification. PCA analysis was performed on the first 100 principal components and plotted using Scanpy.

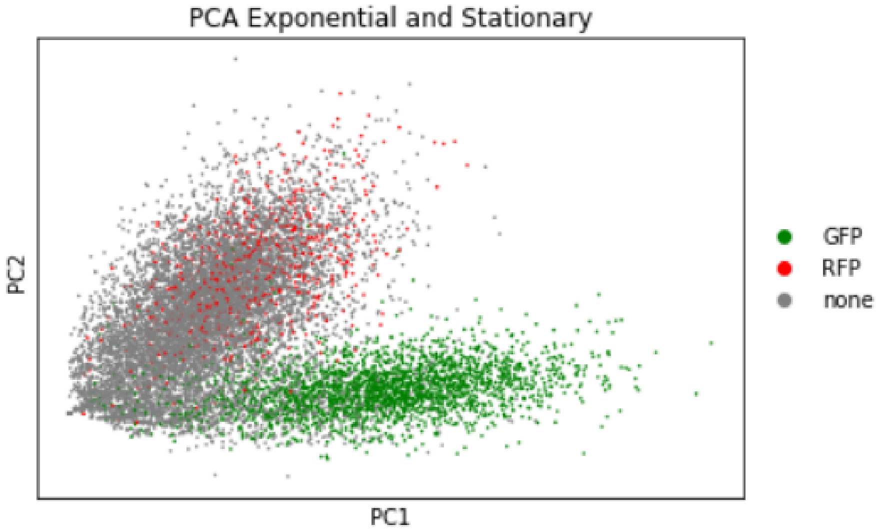


Figure 1. Principal component analysis of Blattman data for gram negative *E. coli* strain MW1655; cells are identified by the presence of GFP (exponential phase) and RFP (stationary phase; none indicates that neither RFP nor GFP plasmids were identified in gene expression

In Figure 1, the principal component analysis confirmed a clear distinction between the exponential and stationary phases of growth demonstrated by the differentiation of cell clustering along the principal components for GFP and RFP expressing cells. Similar to the Blattman PCA results, a majority of cells that did not express either GFP or RFP (none) demonstrated a very high association with stationary cells confirming the results in Blattman et. al (2020).

2D embeddings including t-SNE and UMAP were used to identify cell clusters using the Leiden clustering algorithm provided by the Scanpy toolkit, which is an unsupervised expectation-maximization clustering algorithm that can infer similar cell types without defining the number of clusters. Seen in Figure 2, ,the Leiden algorithm successfully identified distinct clusters for exponential and stationary cells while simultaneously co-associating cells without plasmid to the stationary cells further confirming the conclusion that these cells were in the stationary phase.

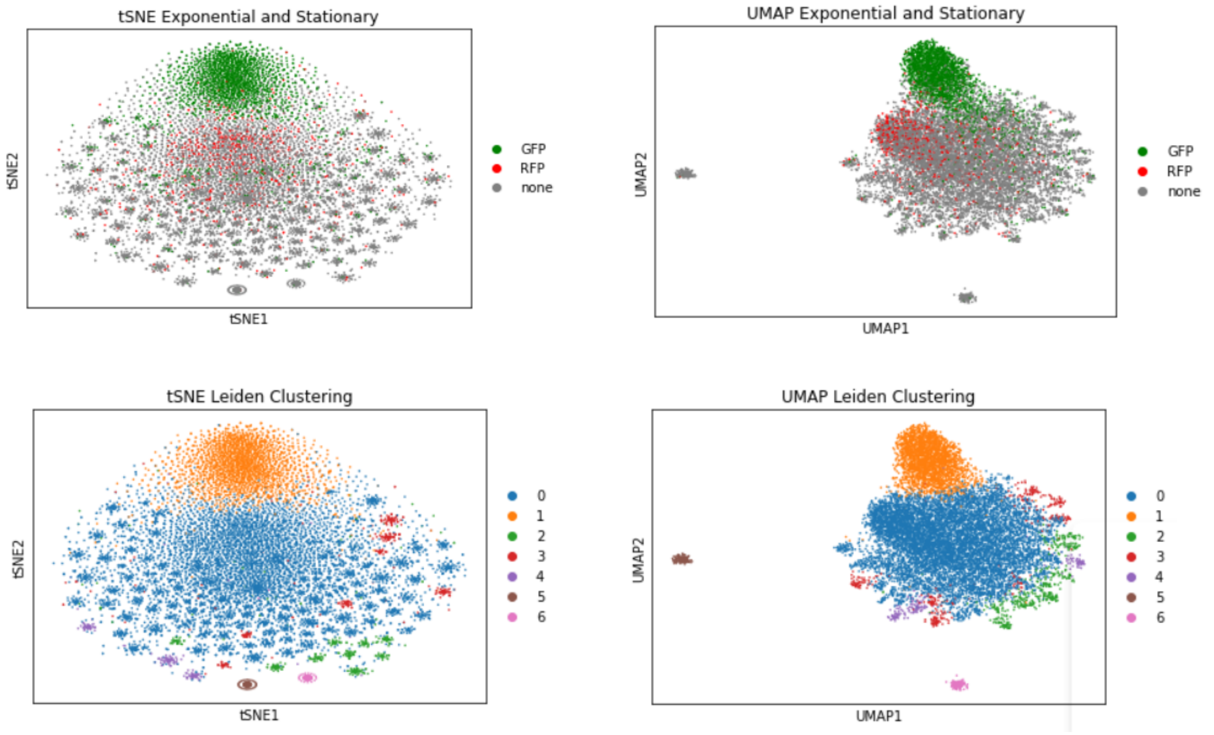


Figure 2. 2D embeddings of Blattman *E. coli* dataset using tSNE and UMAP: tSNE identifying exponential, stationary, and no plasmid cells (top-left); tSNE Leiden clustering(bottom-left); UMAP identifying exponential, stationary, and no plasmid cells (top-right); UMAP Leiden clustering (bottom-right); cells are identified by the presence of GFP (exponential phase) and RFP (stationary phase; none indicates that neither RFP nor GFP plasmids were identified in gene expression

Batch integration of Bacillus subtilis growth curve data shows distinctions in characteristics between algorithms

As preliminary exploration on the capabilities of batch correcting algorithms for bacterial transcriptome data at the single cell level, Harmony, MNN Correct, and ComBat were applied to integrate two separate datasets provided by Kuchina with a total of 25241 cells obtained from Bacillus subtilis sampled at ten optical density (OD) points along the growth curve of the PY79 strain. Both datasets contain cells sampled at different points ranging from OD 0.5 (early exponential phase) to OD6.0 (early stationary phase), with replicates of cells from OD 0.5, 1.0, 1.7, and 2.8.

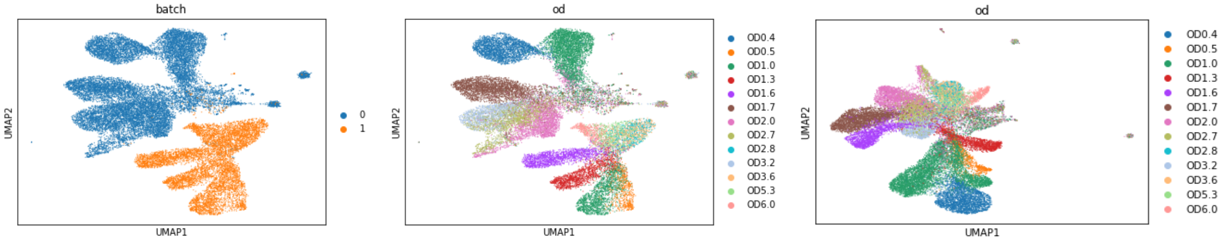


Figure 3. UMAP visualizations of and comBat batch-correction results colored by batch (left) or OD (middle) and MNN batch-corrected data colored by OD (right), combining growth curve data of the Bacillus subtilis PY79 laboratory strain with a total of 25803 cells and 3666 genes sequenced with MicroSPLiT.

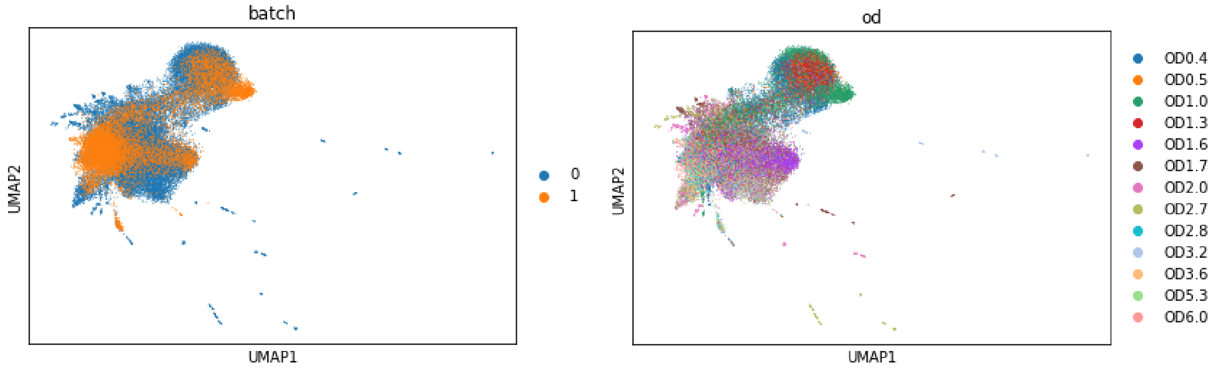


Figure 4. UMAP visualizations of Harmony batch-correction results colored by batch (left) or OD (right).

First, we replicated Kuchina's batch correction method using a Python implementation of ComBat offered by Scanpy. We failed to obtain results reported by Kuchina, as similar OD points (e.g. OD0.4 and 0.5) were not integrated across datasets. This may be due to differences in data labeling, as OD annotations provided by Kuchina differed from their method discussions. To evaluate if different algorithms are better suitable for integration while maintaining temporal ordering, Harmony and MNN Correct were implemented to test batch mixing efficacy.

Visually inspecting a UMAP embedding of the Harmony combined growth curve data colored by optical density (OD), we noticed that different developmental stages were merged, which can be explained by Harmony's usage of fuzzy clustering in which cells are assigned to multiple clusters. While this lack of differentiation could represent the continuous structure of the developmental cells rather than erroneously clustering cells into discrete groups, our visualizations indicate that Harmony simultaneously overcorrected for the effects of discrete batch and continuous cell cycle state factors, resulting in a mixture of batches without preservation of cell type purity, which is not suitable in terms of dataset integration. In comparison, MNN Correct maintained smooth transitions between OD points without incorrectly mixing distinct progenitor populations. However, it should be noted that MNN has a much higher computational runtime and stricter memory requirements, as the distance computation is conducted in the gene expression space.

MNN batch correction of Escherichia coli data unsuitable for integration across different laboratory strains

Next, we applied MNN Correct on *Escherichia coli* sample-by-gene count matrices of 2237 × 914 and 12260 × 914 sequenced by Kuchina et. al (2021) (micro-SPLiT) and Blattman et al. (PETRI-seq) respectively. However, visualizations plots show that there was minimal batch mixing. Potential reasons include lack of sufficient scale-normalization across datasets to account for unbalanced batch size factors and unbalanced repartition of the phenotypes of interest. Additionally, as MNN calculates pairs which represent cells of the same cell type or state from different batches, combining scRNA data sequenced from the different *E. coli* laboratory strains MG1655 (wild-type) and MG1255 (a derivative of MG1655) may have violated the underlying assumption that batch effects are orthogonal to the biological manifold. While further attempts could be conducted to improve MNN results by parameter adjustment and altering data preprocessing steps, other inevitable issues such as long runtimes and loss of the data's mean-variance relationship consequently making MNN-corrected values unsuitable for differential expression analyses disputed the effectiveness of MNN for the purposes of this investigation.

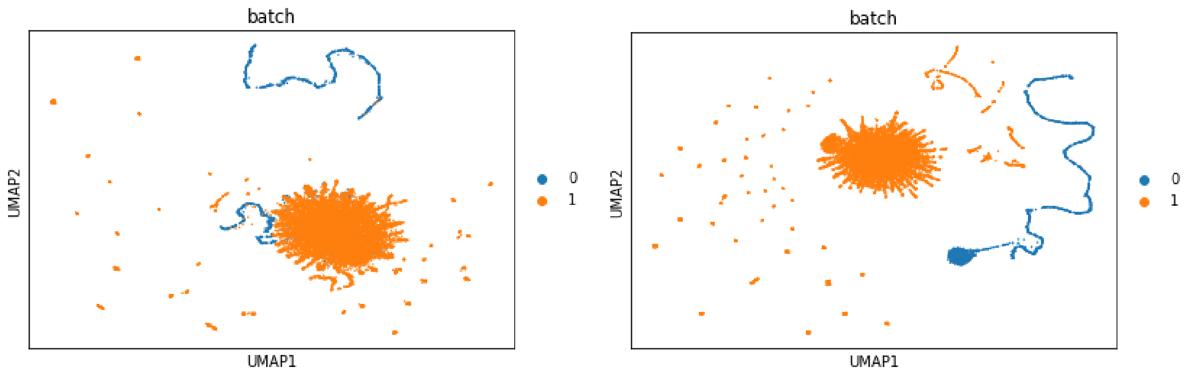
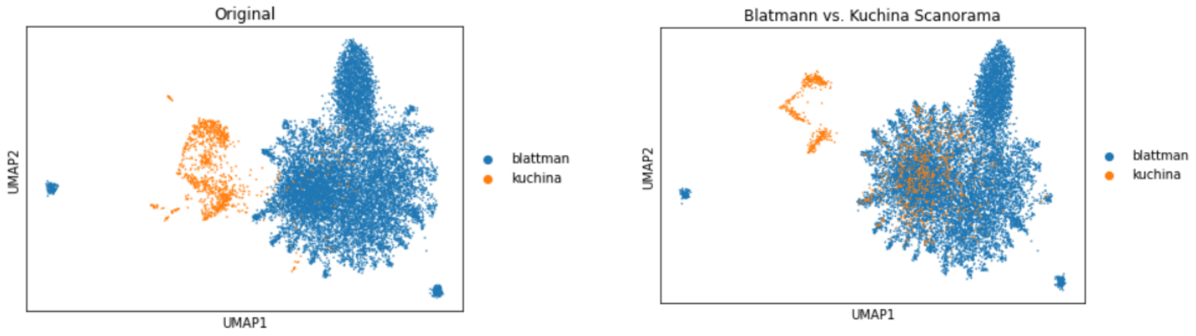


Figure 5. UMAP visualizations of concatenated (left) and batch-corrected (right) using MNN Correct on Escherichia coli scRNA data sequenced by Kuchina et al. (2021) and Blattman et al. (2020), colored by batch. Batch 0 - Kuchina, 1 - Blattman.

Scanorama batch correction produces undesirable integration results

Following MNN, in order to be able to investigate diagonal integration of the Blattman and Kuchina datasets by using Scanorama. For Scanorama, the Blattman and Kuchina datasets were filtered, normalized, log-transformed, and appended with missing genes as zeroes creating zero-inflated datasets. PCA was performed prior to batch integration with Scanorama. Scanorama outputs a low-dimensional representation of the gene expression profiles of the integrated Blattman and Kuchina datasets. Leiden clustering was then performed on the integrated Scanorama low-dimensional embedding to identify distinct cell clusters in the dataset (Figure 6).



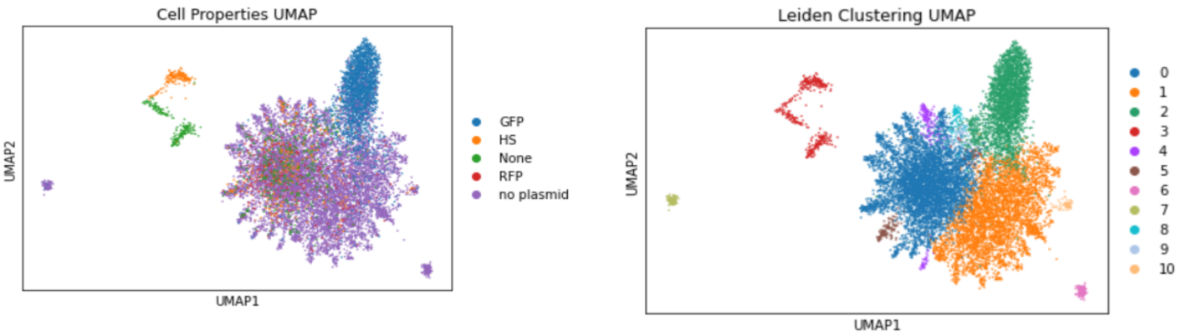


Figure 6. 2D embeddings of uncorrected datasets and Integrated Scanorma *E. coli* dataset using tSNE and UMAP: UMAP using uncorrected datasets from *E. coli* cells in Blattman and Kuchina datasets (top-left); UMAP using Integrated Scanorma *E. coli* dataset (top-right); UMAP identifying exponential (GFP), stationary (RFP), no plasmid cells (no plasmid), heat-shock cells (HS), and non-heat shock cells (None) (bottom-left); UMAP Leiden clustering (bottom-right)

Differential gene expression analysis was performed on Leiden clusters in order to identify important gene markers for analysis and subsequent trajectory inference using Wishbone. The dotplot in Figure 7 demonstrates the top 5 DEG in each cluster using the Wilcoxon rank-sum method in order to identify DEGs by ranking gene expression.

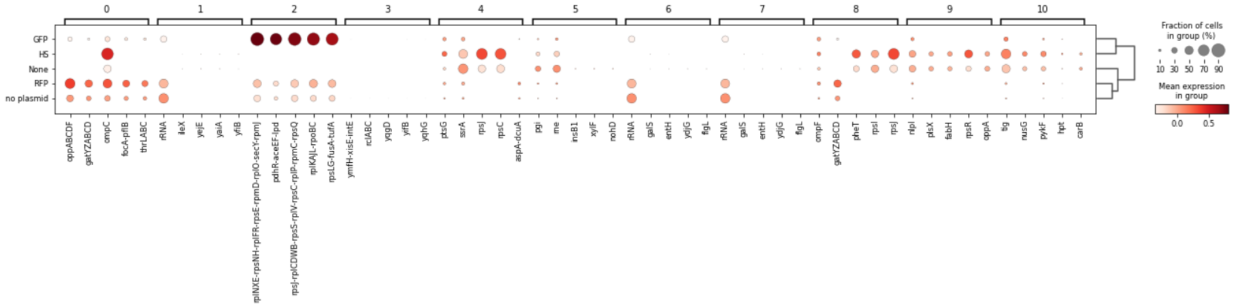


Figure 7. DEG analysis dotplot for top 5 DEGs for each Leiden cluster in association with property identifiers for *E. coli* cells

However, after manual identification of gene functions and associations in the *E. coli* organism, results regarding DEGs were inconclusive and did not demonstrate novel insights into the interaction between heat shock and variable growth stage cells. Trajectory inference using Wishbone was attempted following DEG analysis, however, a clear marker trajectory from the algorithm could not be identified and was unable to be elucidated through use of the Wishbone algorithm. Wishbone trajectories identified by the algorithm can be seen in Figure 8.

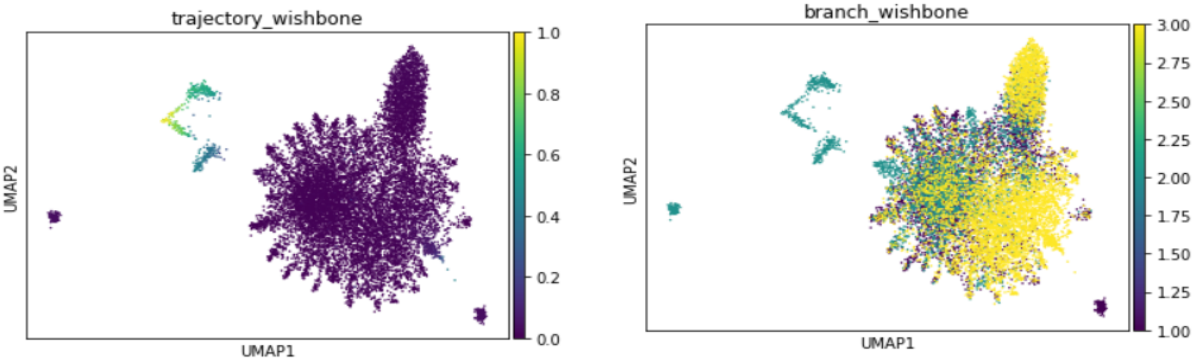


Figure 8. Wishbone-generated bifurcating developmental trajectories on Scanorama results. UMAP representation of the data with each cell colored by trajectory (left) and branch associations (right)

Thus, Scanorama did not prove to be an effective method for batch integration and correction and we decided to use a method that incorporates imputation to resolve missing genes between the two datasets.

scVI imputation and batch-correction shows heterogeneity among heat-shocked cells

For MicroSPLiT and PETRi-seq methodologies, bacterial scRNA sequencing was reported to have low capture rates, as noncoding transcripts accounted for over 90% of sequencing reads in both studies. Therefore, single-cell variational inference (scVI), which is based on a hierarchical Bayesian model using stochastic optimization and deep neural networks for probabilistic representation and analysis of gene expression, was expected to produce better results by accounting for these transcriptional noises as well as batch effects. In addition, scVI imputes observed zeros using autoencoder methods while accounting for batch effects and provides the reconstructed expression matrix for downstream analyses. Since we cross-appended genes as non-expressed to both datasets in order to align variables for vertical integration, we applied scVI to our zero-inflated datasets to find non-biological signals and ameliorate the undersampling of transcript counts in bacterial cells.

After conducting gene augmentation by appending matrices of zero values to both datasets, raw count matrices were concatenated and fit to a scVI model, then trained for 300 epochs. Minimal increase in efficacy after 400 epochs were observed. Using the latent space of the imputed count matrix, we conducted Louvain clustering and found that most clusters contained a mixture of heat shock as well as GFP and RFP expressing cells, which indicates good harmonization. Although *E. coli* MW1255 from the Kuchina dataset was sequenced at OD=0.5 (exponential growth stage), there was a potential differentiation of growth stages indicated by Figure 9 (bottom-right). As microarray studies have suggested that bacteria could form heterogeneous populations as part of a bet-hedging strategy allowing adaptation to heat-shock, we continued to investigate [11].

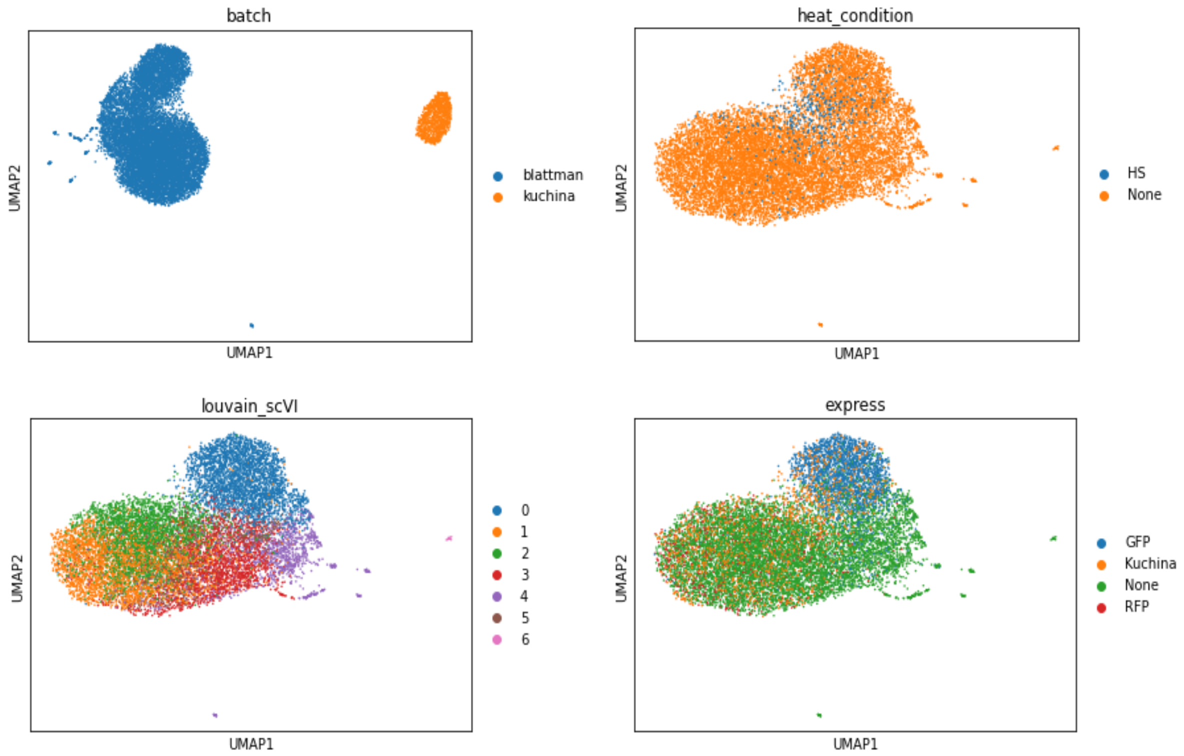
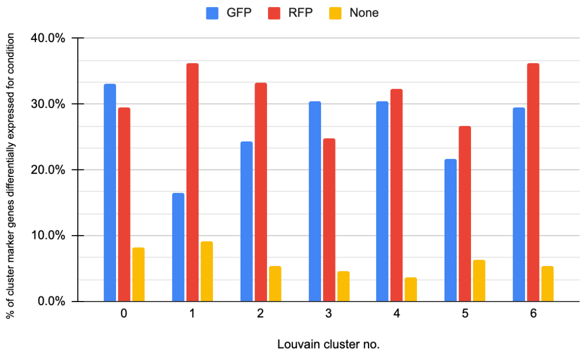
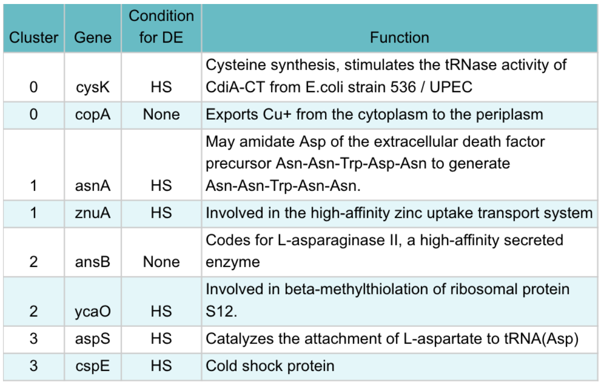


Figure 9. UMAP visualizations of concatenated (top-left) and latent space of scVI imputed count-matrix colored by heat shock condition (top-right), Louvain cluster labels (bottom-left), and expression of fluorescent proteins indicating growth stage (bottom-right).

Since there are no existing cell-type annotation methods for cluster identification on prokaryotic single-cell data in addition to a lack of documentation on growth-stage specific genes, especially as <3% of Escherichia coli genes are known to play roles specific to the stationary phase [12], we conducted manual annotation by computing the proportion of top 200 differentially expressed genes by cluster identified by the function differential_expression_by_cluster from the Scprep package that matched marker genes identified for the original heat-shock, GFP and RFP data (n=100 each) using the function rank_gene_by_group implemented by Scanpy. Results found significantly less GFP (exponential growth) gene markers for cluster 1, 2 and 6, whereas cluster 0 and 3 favored a greater GFP marker gene presence, matching with visual comparisons from UMAP embeddings colored by GFP/RFP expression and Louvain clusters (Figure 10).





Cluster	Gene	Condition for DE	Function
0	cysK	HS	Cysteine synthesis, stimulates the tRNase activity of CdiA-CT from E.coli strain 536 / UPEC
0	copA	None	Exports Cu+ from the cytoplasm to the periplasm
1	asnA	HS	May amidate Asp of the extracellular death factor precursor Asn-Asn-Trp-Asp-Asn to generate Asn-Asn-Trp-Asn-Asn.
1	znuA	HS	Involved in the high-affinity zinc uptake transport system
2	ansB	None	Codes for L-asparaginase II, a high-affinity secreted enzyme
2	ycaO	HS	Involved in beta-methylthiolation of ribosomal protein S12.
3	aspS	HS	Catalyzes the attachment of L-aspartate to tRNA(Asp)
3	cspE	HS	Cold shock protein

Figure 10. Bar chart showing proportion of fluorescent protein expression-related marker genes existing in top DEGs by cluster (left). Table of representative heat condition related genes found specific to clusters 0-3 and main functions (right).

Next, we conducted gene-ontology enrichment analysis with ShinyGO [13], an web-interactive tool able to calculate fold-change values compared to the Escherichia coli str. K-12 substr. MG1655 genome background and visualize overlapping relationships among enriched gene-sets. Upon inspection of the 64 highest differentially genes for *E. coli* cells expressing RFP, processes related with the methionine metabolic and biosynthesis pathways were found to significantly fall below the established FDR p-value threshold and have the highest enrichment scores. Meanwhile, GFP marker genes did not yield enriched methionine-related pathways, instead highlighting iron and other metal ion transport activities potentially indicating high iron uptake by *E. coli* for exponential growth support. These results point to the possibility that a subpopulation of heat-shocked *E. coli* cells heterogeneously bet-hedged into a growth-arrested stationary phase by regulation of methionine biosynthesis. However, this proposition required further supporting evidence, as pathways enriched in heat-shock marker genes did not show similarity to results for GFP or RFP (Figure 11).

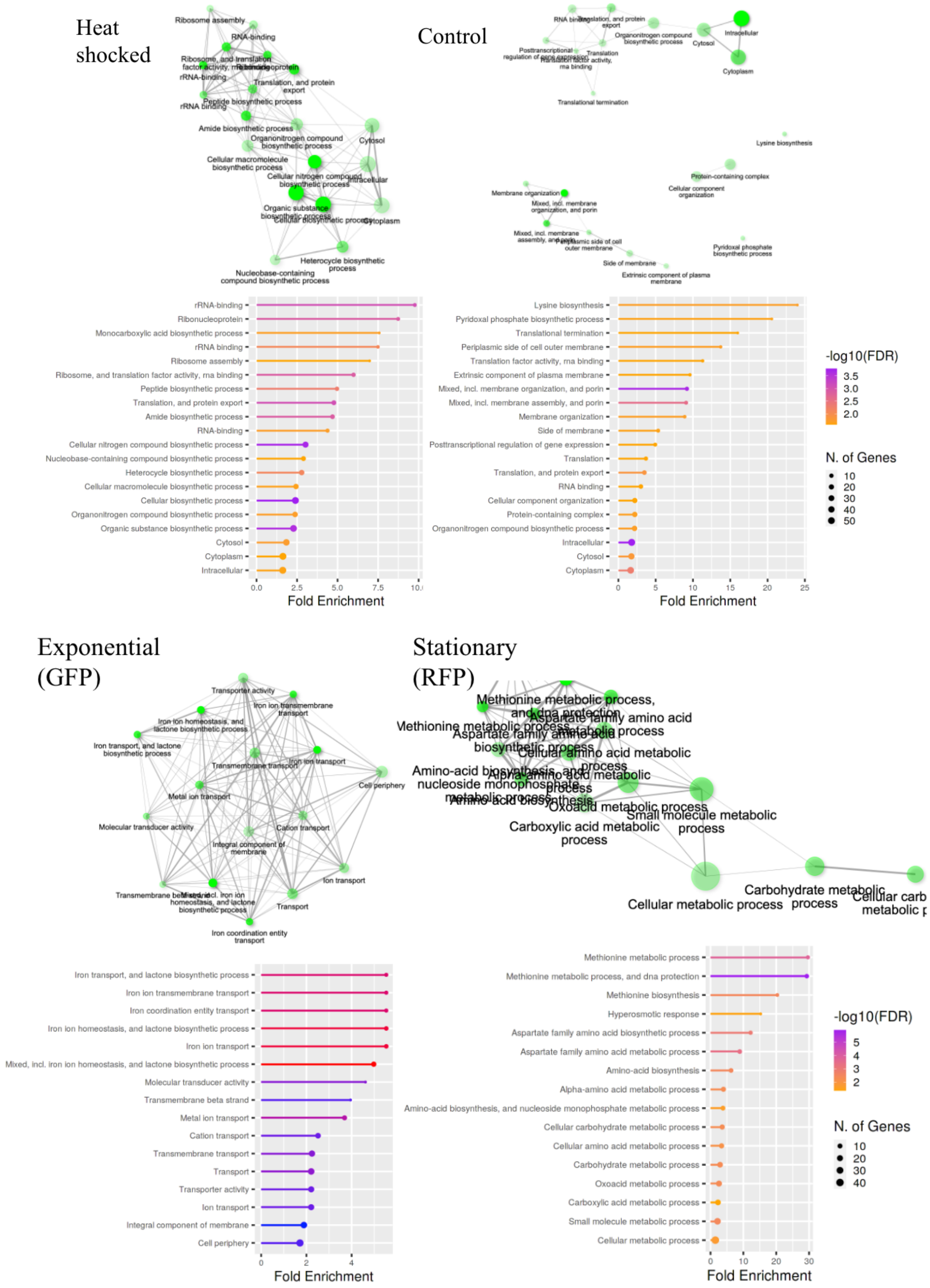


Figure 11. ShinyGO gene-ontology enrichment analysis results for exponential (left) and stationary (right) marker genes identified using scVI-imputed data. Network between pathways (top row), chart of pathways sorted by fold enrichment (bottom row).

To obtain a finer-grained picture, we applied Wishbone [14] to delineate the developmental trajectories for heat-shocked *E. coli* using pseudotime analysis on the scVI latent space for heat-shocked cells. While there have been no conclusive bacterial studies using pseudotime analysis to date, we anticipated that unlike RNA velocity, which requires adaptation of RNA splicing methods for bacteria, Wishbone only requires prior knowledge of a single branch point and could conduct sufficient transition analysis of expression profiles even for prokaryotic scRNA data. The cell with the highest number of expressed genes was selected as the input ‘early cell’.

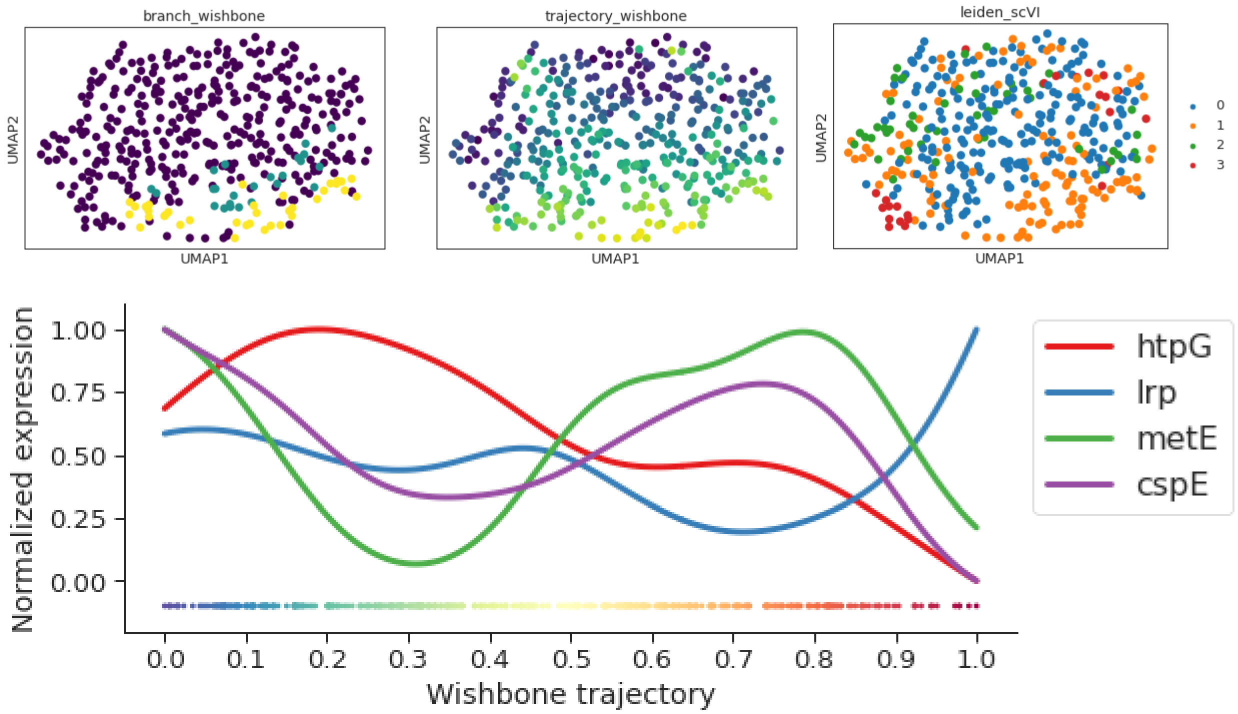


Figure 12. Wishbone-generated bifurcating developmental trajectories on HS from scVI results. UMAP representation of the data with each cell colored by trajectory (top-left), branch associations (top-middle), and Leiden cluster labels (top-right). Plot of marker trends along Wishbone trajectory (bottom). Markers include heat-shock gene htpG, stationary-phase specific gene lrp, methionine biosynthesis enzyme-coding gene metE, and cold-shock gene cspE.

Wishbone recovered a trajectory starting from regions largely assigned to cluster 0 towards 2 and finally 1, with a small grouping of cluster 3 cells toward the end of the progression, likely indicating the subpopulation of cold-shocked genes also identified by Kuchina et. al (2021) to be likely an artifact of a brief cold centrifugation step performed in sample preparation. Combined with postulations from our previous findings in Figure 10, a temporal ordering was found starting from exponential growth followed by gradual entry into dormancy, from which two branches emerged. This hypothesis is further supported by marker trends along the resulting trajectory depicted in Figure 12 (bottom), showing that the pseudo-time point of highest normalized expression for heat shock gene htpG precedes Irp, a dual transcriptional regulator known to affect three-fourths of genes induced upon entry into stationary phase [12]. Additionally, the methionine synthase gene metE showed an elevated level of expression immediately before the lrp-marked growth stage. This finding was surprising as metE is currently understood as an oxidative stress-sensor, while previous literature had established another enzyme in the methionine biosynthesis pathway (metA), to play a major role in induction of a growth-arrested state in response to heat stress [15]. While both datasets showed no significant metA expression across cells likely attributable to technical dropouts, our results add build upon these previous findings to suggest that both first and last step of the methionine biosynthesis pathway, each respectively catalyzed by metA and metE, could act as bottlenecks limiting further growth.

Discussion

Following the investigation of batch correction algorithms using MNN, Scanorama, and scVI to integrate the *E. coli* datasets, scVI proved to be the most successful-method for imputation and harmonizing the two datasets. scVI resulted in a good mixture of cells between both batches that preserved the underlying biological structure of the heat shock conditions and variable growth stages. Most importantly, differences in expression levels for heat-shock response genes in the exponential and stationary phases were significant and may present a form of bet-hedging that caused some cells to enter into a growth-arrested phase earlier than other cells. *E. coli* cells subjected to heat shock conditions demonstrated upregulation of both cold and heat shock proteins in the exponential growth phase and enrichment of genes in the methionine metabolic pathway in the stationary phase. These results indicate that cells exposed to heat shock do not constitutively utilize a common survival strategy and may alter metabolic pathways in multiple ways in response to stress which could suggest a survival tactic by diversifying survival mechanisms. Although imputation violates measurement independence and is still disputed as a reliable method of preserving the original biological structure, the visualizations themselves showed that scVI imputation and batch correction allowed for a more complex representation of the relationship between stress response and bacterial growth than the original isolated Kuchina or Blattman datasets.

A few challenges included finding consistent gene names for both datasets due to incorporation of locus tags in the Kuchina dataset and operon grouping in Blattman, which reduced the inherent number of overlapping genes making vertical integration methods inefficient. Furthermore, there is a lack of available methodologies in the current literature for prokaryotic scRNA-seq which made interpretation of results very difficult and the stochastic heterogeneity in bacterial cells made establishing cell states challenging as the biological distance between cells are minimal leading to a continuum of gene expression profiles rather than biologically distinct and well-separated clusters.

With these challenges in mind, future directions include creating a unified gene reference database for gene annotations for prokaryotic organisms and establishing more information on baseline preprocessing and downstream analysis for prokaryotic scRNA-seq data. Ultimately, new single cell algorithms and tools may need to be developed that specifically account for biological assumptions that differentiate eukaryotic scRNA-seq from prokaryotic scRNA-seq to be able to better investigate and elucidate novel biological insights from prokaryotic scRNA-seq data.

Contributions

All authors conceived and supervised the project with the help of Dr. Elham Azizi, Joy Fan and Cameron Park. Rosenfeld conducted external communications for research guidance, preprocessing and preliminary exploration of Blattman dataset, and batch-integration of Escherichia coli datasets using Scanorama, as well as downstream analyses including visualization, clustering, differential gene expression analysis, and trajectory inference. Li performed preprocessing, gene annotation, and preliminary exploration of Kuchina dataset; batch-integration of Bacillus subtilis datasets, batch-integration and imputation of *E. coli* datasets using MNN Correct and scVI, as well as downstream analyses including differential gene expression analysis, manual cluster identification, trajectory inference, and gene-ontology enrichment analysis. All authors wrote, read and approved the final manuscript.

References

[1] V. Bettenworth et al. Phenotypic Heterogeneity in Bacterial Quorum Sensing Systems. Journal of Molecular Biology, vol. 431, no. 23, pp. 4530-4546 (2019)

[2] Brennan, M. & Rosenthal, A. Single-Cell RNA Sequencing Elucidates the Structure and Organization of Microbial Communities. Frontiers in Microbiology 12, (2021).

[3] Blattman, S.B., Jiang, W., Oikonomou, P., and Tavazoie, S. Prokaryotic single-cell RNA sequencing by in situ combinatorial indexing. Nat Microbiol 5: 1192– 1201. (2020)

[4] Kuchina, A., Brettner, L.M., Paleologu, L., Roco, C.M., Rosenberg, A.B., Carignano, A., et al. Microbial single-cell RNA sequencing by split-pool barcoding. Science 371. (2020)

[5] Imdahl, F., Vafadarnejad, E., Homberger, C. et al. Single-cell RNA-sequencing reports growth-condition-specific global transcriptomes of individual bacteria. Nat Microbiol 5, 1202–1206 (2020)

[6] McNulty, R., Sritharan, D., Liu, S., Hormoz, S., and Rosenthal, A. Z. Droplet-based single cell RNA sequencing of bacteria identifies known and previously unseen cellular states. bioRxiv: 2021.2003.2010.434868. (2021)

[7] A. Ma, et al. Integrative methods and practical challenges for single-cell multiomics. Trends Biotechnol., 38, 1007-1022 (2020)

[8] Haghverdi, L., Lun, A., Morgan, M. et al. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 36, 421–427 (2018).

[9] Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat Biotechnol 37, 685–691 (2019).

[10] Lopez, R., Regier, J., Cole, M.B. et al. Deep generative modeling for single-cell transcriptomics. Nat Methods 15, 1053–1058 (2018).

[11] Kim, S., Kim, Y., Suh, D.H. et al. Heat-responsive and time-resolved transcriptome and metabolome analyses of Escherichia coli uncover thermo-tolerant mechanisms. Sci Rep 10, 17715 (2020).

[12] Tani, T., Khodursky, A., Blumenthal, R., Brown, P. & Matthews, R. Adaptation to famine: A family of stationary-phase genes revealed by microarray analysis. Proceedings of the National Academy of Sciences 99, 13471-13476 (2002).

[13] Ge, S. X., Jung, D., & Yao, R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics, 36(8), 2628–2629.(2020).

[14] Setty, Manu et al. "Wishbone identifies bifurcating developmental trajectories from single-cell data." Nature biotechnology vol. 34,6 (2016).

[15] Schink, S. et al. MetA is a 'thermal fuse' that arrests growth and protects Escherichia coli at elevated temperatures. biorxiv (2021)