Pre-Phaser: Precise Cell-Cycle Phase Detector for scRNA-seq

Alisa Yurovsky
Stony Brook University
ayurovsky@cs.stonybrook.edu

Bruce Futcher
Stony Brook University
bfutcher@gmail.com

Steven Skiena
Stony Brook University
skiena@cs.stonybrook.edu

ABSTRACT

Precise cell cycle phase identification in scRNA-seq improves differential expression analysis. However, the sparseness of scRNA-seq data contributes to the challenge of combining external cell cycle information with gene expression for precise cell phase identification. Existing techniques select their own set of marker genes for each coarse cell cycle phase, and assign cells with enriched average marker expression to this coarse phase.

We observe that precise points along the cell cycle are associated with time course points from microarray experiments that identify cycling genes. In this work, we present the effectiveness of using k-nearest neighbors (kNN) to predict coarse cell cycle phase, and extend the kNN methodology to present the first method to identify precise cell cycle phases by correlating single cell transcript counts for significant cycling genes with time course points.

We demonstrate that tuned kNN outperforms existing methods in assigning coarse cell cycle phase, getting an average F1 score of 0.641 on four previously published scRNA-seq datasets. We then describe *Pre-Phaser*, which establishes a general computational approach for precise cell phase assignment using kNN. Our k-fold cross-validation has an accuracy of 0.872 in precise prediction, and conversion to coarse phase yields an F1 score of 0.716. Our results motivate research into precise cell phase assignment to enable fine-tuned differential expression analyses.

CCS CONCEPTS

• **Applied computing** → **Computational biology**; • **Computing methodologies** → *Machine learning algorithms*.

KEYWORDS

Single Cell, RNA-seq, Nearest Neighbor Classification, Cell Cycle

ACM Reference Format:

Alisa Yurovsky, Bruce Futcher, and Steven Skiena. 2019. Pre-Phaser: Precise Cell-Cycle Phase Detector for scRNA-seq. In *10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB '19)*, September 7–10, 2019, Niagara Falls, NY, USA. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3307339.3342174>

1 INTRODUCTION

Precise cell cycle phase identification in single cell RNA sequencing (scRNA-seq) improves differential expression analysis. The rapidly

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM-BCB '19, September 7–10, 2019, Niagara Falls, NY, USA

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6666-3/19/09...\$15.00

<https://doi.org/10.1145/3307339.3342174>

expanding field of scRNA-seq analysis allows for a greater resolution than bulk RNA-seq [2], enabling studies of custom-defined cell sub-populations in a tissue sample [3]. Differential expression studies with scRNA-seq include cancer studies, inference of cell lineage differentiation trajectories, expression bimodality, etc. [1][3][4][5].

Some scRNA-seq studies regress out cell cycle effects, but others use phase to cluster the cells and explore the expression profiles during different stages of the cell cycle. Precise cell phase identification can stratify scRNA-seq experiments with thousands of cells into very fine consecutive phases and enable fine-tuned studies of cell expression as the cells transition through the cell cycle.

The sparseness of scRNA-seq data contributes to the challenges of combining external cell cycle information with gene expression for precise cell phase identification. Low read counts and high dropout rates make the accuracy of any gene's expression suspect, and multiple genes must be assessed. Another challenge is selecting the external information to use in assigning the cell cycle phase.

Current techniques select their own sets of marker genes for coarse cell cycle phases from literature, and assign cells with enriched average marker expression to a particular coarse phase. Existing programs have two crucial commonalities. First, for each marker gene, they use a single associated label - the coarse cell cycle phase in which this gene has peak expression. And second, they assign each assayed cell to a consensus coarse cell cycle phase.

We observe that precise points along the cell cycle phase are associated with time course points from microarray experiments which identify cycling genes. Studies in yeast [6] and human cells [7][8] have used Fourier analysis to assign a periodicity score to each gene - how strongly the changes in a gene's expression along a time course fit a sinusoidal curve. As a side effect of this analysis we know the precise cell cycle position - the angle between 0 and 2π - for every sampled point in the microarray time course.

We used this insight to develop *Pre-Phaser*: a methodology for correlating time course experiment points with scRNA-seq, in order to obtain precise cell phase for each cell. In lieu of synchronized cells time course experiment analyzed with RNA-seq, which does not exist yet, we use the microarray experiment data as a proof of concept. We adapt the k-nearest neighbors (kNN) methodology, where each single cell profile is compared to the k nearest (microarray) time course data series points, with proper data transformation. Prior work with coarse cell phase assignment, such as Seurat [9] and Cyclone [10], use either simple comparisons, or other machine learning techniques. Our main research contributions are:

- Demonstrating the effectiveness of using kNN for predicting coarse cell cycle phase.
- Extending the kNN methodology to present the first method for precise cell cycle phase identification for single cells.
- Demonstrating the paradox of relative cell phase analysis.

Pre-Phaser algorithms and instructions can be accessed in Github: <https://github.com/ayurovsky/Pre-Phaser>.

Figure 1: Panel A shows phase assignments from the Seurat vignette for murine hematopoietic cells [1]. We used Seurat’s phase assignments to separate the cells into three sub-sets: G1, S, and G2M. Panels B, C, and D show Seurat’s phase assignment on each of the three subsets, where each subset is incorrectly classified into three distinct cell cycle phases.

2 PROBLEMS WITH THE STATE OF THE ART

Multiple studies explore how gene expression of different cell subtypes change with cell cycle phase [1][3][4][5]. Characteristic of the current state of the art, they use marker genes for rough cell cycle stages, cluster cells by strength of expression for a given stage, followed by differential expression analysis. These works motivate both the usefulness of cell phase identification, and the need for a precise scoring strategy.

Seurat [9], one of the leading scRNA-seq processing tools, computes cell cycle scores and assigns cell phases. After normalization and scaling, Seurat uses the phase scoring strategy and marker genes from [11]. S phase score for a cell is the difference of the average normalized expression of the S-marker genes and that of control genes. G2M score is calculated in the same way; if both G2M and S scores are negative, a cell is classified as G1 phase, otherwise its phase is determined by the maximum of the G2M and S scores.

We find two limitations of Seurat methodology. First, while classification of cells into G1, S, and G2M phases captures coarse-grained distinctions, the boundaries separating the cell cycle phases are somewhat blurry, since gene expression profiles peak at different points spanning each phase. The cell cycle scores calculated by Seurat are a proxy for the strength of evidence of the cell being in a given phase, but do not clarify its more precise location.

The second problem with the Seurat lies in the fact that its phase assignment assumes that the cells come from a spectrum of cell phases and will lead to misclassification if the cells are fully or partially synchronized. In Figure 1, panel A, we see the phase assignments from the Seurat vignette [1]. We used Seurat’s phase assignments to separate the data into three sub-sets: G1, S, and

G2M. Figure 1 panels B, C, and D show the results of Seurat’s phase assignment on the three subsets. Not surprisingly, each subset is again classified into three distinct cell cycle phases, which cannot be true. We note that Seurat’s goal of regressing out the cell cycle effect using phase scores is unaffected by this misclassification.

Figure 2: Distribution of 1230 human cell cycle regulated genes from a microarray time course experiment [8]. The angle of each gene represents the time of its peak expression in the cell cycle. Color represents the coarse phase of each gene’s peak expression. Color coherence demonstrates the correspondence between angles and phases.

Figure 3: Recursive Feature Selection for coarse phase prediction using kNNs on the McDavid_Train Training Dataset.

It is clear that we need a new methodology to place cells more precisely along the cell cycle, independent of the underlying cell phase distribution of the sample.

3 IDEA: PRECISE CELL PHASE ASSIGNMENT

We observe that precise points along the cell cycle phases are associated with time course points from microarray experiments which identify cycling genes. The cell cycle is divided into three coarse stages: G1, S, and G2M. While housekeeping genes maintain stable expression in cycling cells, other genes peak in different points along the cycle. Many microarray cell cycle studies in yeast [6] and human cells [7][8] have used Fourier analysis to assign the periodicity score for each gene, measuring how strongly the changes in a gene's expression along the time course fit a sinusoidal curve.

Figure 2 shows phases of cell cycle regulated expression, modeled on supplemental figure from Whitfield et. al [7], with data from Grant et. al. [8], both of whom performed microarray experiments on the time course data to identify cell-cycle regulated human genes. The radial position (angle) of each gene represents the time of its peak expression during the cell cycle. We can see that, while there is some overlap, the coarse cell cycle stages are clearly identified by contiguous stretches of corresponding genes with peak expression.

$$A = \sum \sin(2\pi(t/T) + \phi) \log_2(\text{ratio}(t)) \quad (1)$$

$$B = \sum \cos(2\pi(t/T) + \phi) \log_2(\text{ratio}(t)) \quad (2)$$

Briefly, the microarray time course studies have used the sine (A) and cosine (B) coefficients from the Fourier Transform (eqs. 1 and 2) to determine the strength of the periodicity of cell cycle genes, where T is the cell cycle period, t is the time after release, ϕ is the phase offset, and $\text{ratio}(t)$ is the normalized Cy5/Cy3 (synchronized to unsynchronized cells) expression ratio at time t .

As a side effect of the Fourier analysis, we know the period length T and the phase offset ϕ for each microarray time course. Thus we also know the precise cell cycle position - the angle between 0 and 2π - for each point on the microarray time course.

For each cell, we can compare its profile to the best-matching microarray (or other assay such as RNA-seq) time course data series points, to find the precise angle position. Since we want to use several best-matching time points to zero into the precise position, we employ nearest neighbor methods.

4 KNN FOR COARSE PHASE PREDICTION

kNN, an example of a nearest neighbor algorithm, is a non-parametric classification method, particularly suited for our problem because we interested in obtaining the consensus of neighborhood surrounding our test data point. In nearest neighbor classification, the output $\hat{Y}$ is predicted based on the points that are nearest in the training set to query x :

$$\hat{Y}(x) = \frac{1}{k} \sum_{n_i \in N_k(x)} n_i \quad (3)$$

where $N_k(x)$ is the neighborhood of x , defined by k points in training set that are closest to x . The assumption of the nearest neighbor method is that the optimal decision function is well approximated by a locally constant function.

The relative order of the points in a neighborhood is decided by a distance metric. The hyperparameters for kNN are the number (and which) features to use, and the number of nearest neighbors.

4.1 Basic Implementation Details:

We use Euclidean distance as the distance metric for nearest neighbor selection. We identify the labeled phase of each of the k nearest neighbors, and assign the majority class as the classified phase. Ties are broken randomly.

For evaluation, we calculate accuracy as percent of single cells with correctly identified phase from the labeled datasets. We use the F1 score to compare our results with published work:

$$F1_score = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (4)$$

4.2 Test and Training Data

Large scRNA-seq datasets labeled with known cell phase ground truth are surprisingly rare. Table 1 summarizes the labeled datasets that we used to train and test our methodology. All datasets were labeled with three coarse phases (G1, S, G2M), and were reasonably balanced among the three classes. In all cases, we use the intersection of training and test datasets for prediction. Genes with missing data are excluded from the analysis.

Table 1: Datasets Labeled with Coarse Cell Phase

Short Name	Cells	Genes	Assay Type
McDavid_Train [12]	500	250	human scRNA-seq
McDavid_Test [12]	430	250	human scRNA-seq
Sasagawa [13]	35	36,794	mouse Quartz-seq
Dolatabadi [14]	90	82	human scRNA-seq
Buettner [15]	182	34,423	mouse scRNA-seq
Scialdone [10]	3	38,050	mouse bulk RNA-seq

Figure 4: Recursive Feature Selection for precise angle prediction using kNNs on the Grant Training Dataset.

We selected the cell-phase labeled dataset with the largest number of cells that we could find — 930 cells from human cell lines from McDavid, et al. [12] — and split it into a 500 cell McDavid_Train and 430 cell McDavid_Test sets. We used three samples (G1, S, and G2M) from Scialdone dataset to compare algorithms on phase-labeled bulk data, but did not include it in average performance calculation.

4.3 Data Normalization

Data normalization is very important to ensure that the values in the train and test data sets are in the same range. Various normalization strategies have been applied to single cell data before analysis [16].

We employ a simple normalization strategy similar to Seurat. First, we take the natural log of all the values. Second, because we need to compare data from different cell lines produced by different assays, we further normalize each gene (feature) in the dataset by converting it to z-scores (after confirming that the values are approximately normally distributed).

4.4 Optimizing Feature Selection

Our hyperparameters for kNN are the number (and which) features to use, and the number of nearest neighbors to consider. The curse of dimensionality, defined by Bellman [17], implies that in our high dimensional space (each gene expression is a feature), the distance function used to define a local neighborhood loses its usefulness because the nearest data points are very far from each other.

We implemented Recursive Feature Selection, a greedy optimization algorithm. We obtain our kNN method accuracy for each feature separately. We select the best performing feature, and then re-run for all possible sets of two features. Recursive selection of features continues until the full set is reached. The peak for each k represents the set of genes with the best global accuracy.

We used the McDavid_Train dataset for k -fold cross validation. In Figure 3 we can see that for most k , the accuracy peaks at less than 100 features. We select the corresponding feature sets, intersected with test feature sets for evaluation.

Evaluating on our McDavid_Train set for the number of neighbors k in the range [1-30], we find that the maximum test accuracy is

achieved at $k=10$. However, when evaluating on the McDavid_Test, we find that the maximum accuracy is achieved at $k=3$; we use this number of k -nearest neighbors for our evaluation.

4.5 10-Fold Cross-Validation

In order to ensure that we have split the 930 McDavid dataset into training and test sets in an unbiased fashion, we perform 10-fold cross-validation. We make ten random splits of the 930 cells into McDavid_Train and McDavid_Test, and evaluate our method on the McDavid_Test as well as all other test datasets.

The accuracies range in [0.691, 0.760] for McDavid_Test, [0.489, 0.600] for Dolatabadi, [0.686, 0.771] for Sasagawa and [0.527, 0.593] for Buettner. The variance for accuracy is fairly low for all datasets, demonstrating the robustness of our training set choice.

5 PRE-PHASER: PRECISE CELL PHASE PREDICTION METHODOLOGY WITH KNNs

5.1 Training Data

Because we do not have a true RNA-seq time course on time series with validated phase angles, we have to use a microarray experiment with precise angles as our training set.

We use the Grant et. al. [8] dataset, where microarray experiment was performed on the time course time series data to identify periodically expressed cell-cycle regulated human genes. Time series data at two hour intervals was collected from four independent time courses on synchronized human U2OS osteosarcoma-derived cells. Isolated total RNA was used to determine the expression of all genes in the genome using Agilent. Fourier transforms [7] were employed to generate a periodicity score for each gene, and this information was used by us to obtain exact phase angles for each time point. After filtering and data transformation we obtained 1230 cell cycle regulated genes.

5.2 Microarray Data Transformation

In order to make an apples-to-apples comparison between microarray expression values and single cell expression values, we must transform one of the dataset into the other type.

Micro array data are log ratios of expression of synchronized time course cells to an asynchronous population, expressing the degree that a certain gene is over or under expressed. Each time course point is a set of expression values for an order of a thousand cell-cycle regulated genes, and can be thought of as a snapshot at a particular cell cycle angle, analogous to a single cell profile. To transform microarray values into single cell counts, we take each gene's microarray expression out of log space, and then multiply by this gene's median TPM (transcripts per million) obtained from the GTex Analysis Version 7 [18]. We tested all GTex tissue types, and found that evaluating our transformed single cell dataset with Seurat [9], yielded the best accuracy with Skin-SunExposed(Lowerleg) tissue; we use these TPM values for our microarray to single cell dataset transformation.

The reverse direction, taking a single-cell dataset and transforming each single cell into a microarray time-course point, has proven to be much more noisy, and was not employed. We normalize our transformed microarray dataset as all other datasets in Section 4.

Table 2: We show F1 scores for coarse cell phase prediction, comparing the performance of baseline label selections, five competitor algorithms, and our kNN methodology. We note that Cyclone algorithms achieve perfect F1 scores on the Buettner dataset by default, because it is their training dataset.

Algorithm (row), Dataset (col)	McDavid_Test	Dolatabadi	Sasagawa	Buettner	Average
Baseline Random Pick	0.333	0.337	0.331	0.333	0.334
Baseline Top Class Pick	0.185	0.171	0.242	0.175	0.193
Seurat	0.518	0.584	0.444	0.568	0.529
Cyclone PCA	0.410	0.33	0.840	0.700	0.570
Cyclone Random Forest	0.210	0.480	0.600	1.000	0.573
Cyclone LR	0.230	0.460	0.510	1.000	0.550
Cyclone LRAI	0.130	0.430	0.570	1.000	0.533
kNN	0.703	0.536	0.762	0.562	0.641

5.3 kNN Implementation Details

For precise cell phase prediction, we adapt the kNN methodology from Section 4. Once we obtain the k nearest neighbors from the training set, instead of using the majority class label, we take the average of the phase angles associated with each training sample to determine the precise predicted angle:

$$\text{predicted_angle} = \text{atan2}\left(\sum_{n=1}^k \sin \angle_n, \sum_{n=1}^k \cos \angle_n\right) \quad (5)$$

For Evaluation:

For precise cell phase prediction evaluation, the correct measure is accuracy - i.e. how far is the *predicted_angle* from the *real_angle*. Since these angles are located on the circle, the largest positive distance between two angles is π .

$$\text{test_error} = \frac{\pi - ||\text{real_angle} - \text{predicted_angle}| - \pi|}{\pi} \quad (6)$$

$$\text{accuracy} = 1 - \frac{\sum_{n=0}^m \text{test_error}}{m - 1} \quad (7)$$

Equation 6 defines *test_error* as the difference between *real_angle* and *predicted_angle*, as a percentage of the difference range (π). The accuracy in Equation 7 is thus 1 minus the average test_error.

Since the only dataset with precise cell phase information that we have is our microarray training data, we can not use the precise predicted angle for a fair evaluation of our method, and must transform the predicted precise angles into coarse cell phases.

We use the coarse phase assignments for each of the four time courses points in [8] and reconcile them with the corresponding angles for each of the time course points to divide the circle into three sectors.

We use the coarse phase assignments for evaluation same as in Section 4. We retain the precise predicted angles information, as well as the precise accuracy, and use it for the k-fold validation.

5.4 Feature and Neighbor Selection

In the same way as for coarse cell phase prediction in Section 4, we determine most predictive set of features and the best k for the number of neighbors using k-fold validation on the training set.

Figure 4 shows that we obtain the best (0.872) k-fold validation accuracy at $k=10$ and using a small number (78) of genes.

6 RESULTS

We compare our performance on labeled datasets from Table 1 against popular packages with coarse cell phase assignment: Seurat [9] and Cyclone [10]. For Cyclone, we evaluate four of their top techniques in the python demo: PCA, randomForest, LR, and LRAI.

Before evaluation, two small modification were done. For Seurat input, we pad smaller datasets with extra copies of non-marker genes to increase the control set used for random selection. For Cyclone, we found that the classification report was produced on training data and modified it to use test datasets. Default parameters were used for all competitor methods.

All results are shown in Table 2. For clarity, we include the F1 scores of two baseline predictors, one that randomly assigns a class label, and one that always selects the top class. In summary, **our average F1 score (0.641) across four test datasets is higher than the other tested approaches**. For three bulk Scialdone samples, one for each phase, we correctly predict the phase for two out of three samples, same as Seurat, but beating Cyclone. We have shown the effectiveness of using kNN for coarse phase prediction on mouse and human scRNA-seq datasets from a variety of cell types and assays.

For our precise cell phase prediction, we first show the results of k-fold cross validation on the training dataset —our only dataset with precise angles. We demonstrate the accuracy of precise cell phase prediction by plotting real angles vs predicted angles for all time points in Figure 5. We see that most points line up along the diagonal, indicating that **the predicted angle is extremely close to the real angle**, which is our desired outcome. Since we present a linear view of the angles on a circle, well-predicted red-colored S phase time points may appear in both the top and the bottom red-colored rectangles, and are still fairly close to the diagonal. We observe that the G2M (blue) time points are classified most precisely, and the G1 (green) time points least precisely.

Our F1 scores for coarse phase based on precise cell phase prediction, shown in Table 3, are generally low. We have a high 0.716 F1 score on cross-validation (Grant) dataset, demonstrating the potential of our approach. We have same F1 score as the best coarse phase predictors on the Scialdone dataset (correctly predicting the phase for two out of three bulk samples), demonstrating that we do well on bulk sequencing of synchronized cell population, which is the same type as our microarray training set.

Figure 5: Real vs. predicted angles from the k-fold cross validation on the Grant microarray training dataset. Most points line up on the diagonal, demonstrating the accuracy of *Pre-Phaser* precise cell phase prediction.

Table 3: F1 Scores for Precise Angle Prediction, Converted to Coarse Phase for Evaluation

McDavid_Test	Dolatabadi	Sasagawa	Buettner	Scialdone	Grant
0.478	0.332	0.477	0.447	0.571	0.716

7 RELATED WORK

A variety of scRNA-seq processing packages have been developed in recent years. MAST [19] is a statistical framework to characterize heterogeneity in scRNA-seq using a generalized linear model. Numerous approaches order cells in unsupervised manner [20][21][22][23]. SINCERA [24] provides a single-cell RNA-seq processing framework, which includes normalization by trimmed mean and per-genes, clustering and cell type assignment, differential expression, and other analyses. One of the leading scRNA-seq packages called Seurat [9] has been described in Section 2.

Buettner et. al [15] use single-cell latent variable model (scLVM) to group cells by cell cycle. In follow-up work to scLVM, Scialdone et. al. [10] compare a number of supervised machine learning methods to assign cell cycle stage based on single cell data.

Focusing on bimodal cell cycle gene expression, McDavid et. al characterize genes using a semi-continuous modeling framework based on a generalized linear model [12]. In another machine learning approach, Dolatabadi et. al applied a random forest classifier with 1000 trees to identify cell cycle phase for single cells based on the expression of 93 marker genes [14].

A variety of distance metrics can be used in nearest neighbor classification. Hu et. al [25] explored the effects of using different distance metrics on nearest neighbor classification of medical images, comparing Euclidean, cosine, Chi square, and Minkowsky.

A recent review on dimensionality reduction by feature selection [26] grouped techniques into three classes. Fast filter methods, which include t-statistics, Fisher's discriminant ratio, correlation-based filters, etc., ignore the interaction with the classifier and the dependencies between the features. Wrapper methods, which include forward and backward sequential selection and elimination, wrap a model in a search algorithm, suffer from possible overfitting, balanced against taking into account feature dependencies and interaction with the classifier. Embedded methods contain feature elimination or selection as part of the classification process, and include decision trees, random forests, and weighted naive Bayes.

For nearest neighbor classification specifically, there is an assumption that class probabilities are relatively constant in each local neighborhood. In a high-dimensional problem, class probabilities may only change in a small sub-space, which is addressed by discriminant adaptive nearest-neighbor (DANN) method of Hastie and Tibshirani [27].

8 CONCLUSION

We demonstrated the effectiveness of nearest neighbor classification approach for coarse cell phase prediction by comparing our algorithm against five competitor algorithms on four diverse datasets.

We showed how precise cell phase prediction can be based on the time course sequencing experiments. For our precise cell phase methodology, very high performance on cross-validation and same type data source, while low performance on the single cell datasets leads us to two conclusions:

- We have shown that kNN approach is very effective for precise cell phase detection.
- We demonstrate a need for Fourier analysis on a bulk RNA-seq time course with synchronized human or mouse cells.

An important future direction for *Pre-Phaser* is to be able to differentiate between actively cycling cells that are in G1 phase vs. quiescent cells that are in "shallow G1" or G0 phase.

Precise cell phase assignment will enable studies of gene expression in any desired slices of cell phases. This will allow the researchers to zoom into the differential expression at a particular point of a cell cycle from collected tissue cells, without the need for synchronized in vitro culture experiments.

ACKNOWLEDGMENTS

This work was done with support from the NSF grants GRFP-1315232 and IIS-1546113.

REFERENCES

- [1] Sonia Nestorowa, Fiona K. Hamey, Blanca Pijuan Sala, Evangelia Diamanti, Mairi Shepherd, Elisa Laurenti, Nicola K. Wilson, David G. Kent, and Berthold Göttgens. A single cell resolution map of mouse haematopoietic stem and progenitor cell differentiation. *Blood*, 2016.
- [2] Daniel Hebenstreit and Sarah A Teichmann. Analysis and simulation of gene expression profiles in pure and mixed cell populations. *Physical biology*, 8:035013, 06 2011.
- [3] Alex K Shalek, Rahul Satija, Xian Adiconis, Rona S Gertner, Jellert T Gaubomme, Raktima Raychowdhury, Schraga Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, John J Trombetta, Dave Gennert, Andreas Gnirke, Alon Goren, Nir Hacohen, Joshua Z Levin, Hongkun Park, and Aviv Regev. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–240, 06 2013.
- [4] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nature biotechnology*, 32(4):381–386, 04 2014.
- [5] Sean C Bendall, Kara L Davis, El-ad David Amir, Michelle D Tadmor, Erin F Simonds, Tiffany J Chen, Daniel K Shenfeld, Garry P Nolan, and Dana Pe'er. Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, 157(3):714–725, 04 2014.
- [6] P T Spellman, G Sherlock, M Q Zhang, V R Iyer, K Anders, M B Eisen, P O Brown, D Botstein, and B Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular biology of the cell*, 9(12):3273–3297, 12 1998.
- [7] Michael L Whitfield, Gavin Sherlock, Alok J Saldanha, John I Murray, Catherine A Ball, Karen E Alexander, John C Matese, Charles M Perou, Myra M Hurt, Patrick O Brown, and David Botstein. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular biology of the cell*, 13(6):1977–2000, 06 2002.
- [8] Gavin D Grant, 3rd Brooks, Lionel, Xiaoyang Zhang, J Matthew Mahoney, Viktor Martynov, Tammara A Wood, Gavin Sherlock, Chao Cheng, and Michael L Whitfield. Identification of cell cycle-regulated genes periodically expressed in u2os cells and their regulation by foxm1 and e2f transcription factors. *Molecular biology of the cell*, 24(23):3634–3650, 12 2013.
- [9] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36:411 EP –, 04 2018.
- [10] Antonio Scialdone, Kedar N. Natarajan, Luis R. Saraiva, Valentina Proserpio, Sarah A. Teichmann, Oliver Stegle, John C. Marioni, and Florian Buettner. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54 – 61, 2015. Inferring Gene Regulatory Interactions from Quantitative High-Throughput Measurements.
- [11] Itay Tirosh, Benjamin Izar, Sanjay M. Prakadan, Marc H. Wadsworth, Daniel Treacy, John J. Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, Mohammad Fallahi-Sichani, Ken Dutton-Regester, Jia-Ren Lin, Ofir Cohen, Parin Shah, Diana Lu, Alex S. Genshaft, Travis K. Hughes, Carly G. K. Ziegler, Samuel W. Kazer, Aleth Gaillard, Kellie E. Kolb, Alexandra-Chloé Villani, Cory M. Johannessen, Aleksandr Y. Andreev, Eliezer M. Van Allen, Monica Bertagnolli, Peter K. Sorger, Ryan J. Sullivan, Keith T. Flaherty, Dennis T. Frederick, Judit Jané-Valbuena, Charles H. Yoon, Orit Rozenblatt-Rosen, Alex K. Shalek, Aviv Regev, and Levi A. Garraway. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196, 2016.
- [12] Andrew McDavid, Lucas Dennis, Patrick Danaher, Greg Finak, Michael Krouse, Alice Wang, Philippa Webster, Joseph Beechem, and Raphael Gottardo. Modeling bi-modality improves characterization of cell cycle on gene expression in single cells. *PLOS Computational Biology*, 10(7):1–10, 07 2014.
- [13] Yohei Sasagawa, Itoshi Nikaido, Tetsutaro Hayashi, Hiroki Danno, Kenichiro D. Uno, Takeshi Imai, and Hiroki R. Ueda. Quartz-seq: a highly reproducible and sensitive single-cell rna sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biology*, 14(4):3097, 2013.
- [14] Soheila Dolatabadi, Julián Candia, Nina Akrap, Christoffer Vannas, Tajana Tesan Tomic, Wolfgang Losert, Göran Landberg, Pierre Åman, and Anders Ståhlberg. Cell cycle and cell size dependent gene expression reveals distinct subpopulations at single-cell level. *Frontiers in genetics*, 8:1; 1–1, 01 2017.
- [15] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33:155 EP –, 01 2015.
- [16] Davis J McCarthy, Kieran R Campbell, Aaron T L Lun, and Quin F Wills. Scater: pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r. *Bioinformatics*, 33(8):1179–1186, 04 2017.
- [17] Richard E Bellman. *Adaptive control processes: a guided tour*. Princeton University Press, 1961.
- [18] GTEx Consortium. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 06 2013.
- [19] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, Hannah W. Miller, M. Juliana McElrath, Martin Prlc, Peter S. Linsley, and Raphael Gottardo. Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome Biology*, 16(1):278, 2015.
- [20] Manu Setty, Michelle D Tadmor, Shlomit Reich-Zeliger, Omer Angel, Tomer Meir Salame, Pooja Kathail, Kristy Choi, Sean Bendall, Nir Friedman, and Dana Pe'er. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature biotechnology*, 34(6):637–645, 06 2016.
- [21] Charles A Herring, Amrita Banerjee, Eliot T McKinley, Alan J Simmons, Jie Ping, Joseph T Roland, Jeffrey L Franklin, Qi Liu, Michael J Gerdes, Robert J Coffey, and Ken S Lau. Unsupervised trajectory analysis of single-cell rna-seq and imaging data reveals alternative tuft cell origins in the gut. *Cell systems*, 6(1):37–51.e9, 01 2018.
- [22] Rhonda Bacher and Christina Kendzioriski. Design and computational analysis of single-cell rna-sequencing experiments. *Genome Biology*, 17(1):63, 2016.
- [23] Ron C. Anafi, Lauren J. Francey, John B. Hogenesch, and Junhyong Kim. Cyclops reveals human transcriptional rhythms in health and disease. *Proceedings of the National Academy of Sciences*, 2017.
- [24] Minzhe Guo, Hui Wang, S. Steven Potter, Jeffrey A. Whitsett, and Yan Xu. Sincera: A pipeline for single-cell rna-seq profiling analysis. *PLOS Computational Biology*, 11(11):e1004575–, 11 2015.
- [25] Li-Yu Hu, Min-Wei Huang, Shih-Wen Ke, and Chih-Fong Tsai. The distance function effect on k-nearest neighbor classification for medical datasets. *SpringerPlus*, 5(1):1304; 1304–1304, 08 2016.
- [26] Yvan Saeys, IÅsaki Inza, and Pedro LarraÅsaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 08 2007.
- [27] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE Pattern Recognition and Machine Intelligence*, 18:607–616, 1996a.