UNIVERSITY OF CALIFORNIA,
IRVINE

Leveraging Latent Representations to Investigate Biological Processes

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematical, Computational and Systems Biology

by

Michelle Nguyen Ngo

Dissertation Committee:
Professor Babak Shahbaba, Chair
Professor John Lowengrub
Professor Jun Allard

2023

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

LIST OF ALGORITHMS

ACKNOWLEDGMENTS

tauFisher is supported by an award entitled "Skin Biology Resource-Based Center at UCI - Systems Biology Core" from the Center of Complex Biological Systems funded by NIH/NIAMS P30-AR075047 (J.D. and M.N.); National Institute of Health grants R01AR056439 and P30AR075047 (B.A.); NSF grant DMS1763272 and a grant from the Simons Foundation (594598) (J.D., M.N. and B.A.); National Institute of General Medical Sciences, National Research Service Award GM136624 (J.D.); NSF grant DMS1936833 (M.N. and B.S.); and the California Institute for Regenerative Medicine Training Program Award EDUC4-12822 (S.S.K.).

The differential gene, clustering and biclustering projects are supported by the NSF grant DMS1763272, Simons Foundation grant (594598), and NSF grant DMS1936833.

It seems almost criminal to reduce all the support I've received over the course of my academic lifetime to a minor section of my thesis. While this thesis was conceived and completed over the past five or so years, it would be remiss to ignore all the people that shaped my academic path and led me to my current career even if no one reads this.

First and foremost, I'd like to thank my committee members: Babak Shahbaba, John Lowengrub and Jun Allard for their support, encouragement and flexibility. Without their insight and guidance, particularly Babak, I would not have the opportunity to present this body of work. You've all taught me different aspects of perseverance and given me examples of what it means to be a great researcher.

I'd like to thank my collaborators: these projects would not have been fully conceived without your partnership, and thank you for an immensely enjoyable experience. I'd also like to thank the MCSB and CMCF administrative staff for their work behind the scenes. Then going back in time: thank you to Professor Shu-Min Liao, who introduced me to the field of biostatistics. I credit you for the pivotal point in my career, and I'm forever grateful.

Regarding the personal side, I want to first thank my family for their unconditional love and support. Thank you for always providing me with food and also keeping my friends well-fed. I'm especially thankful that being in Irvine has had the added benefit (sometimes drawback) of being close to many of you. Me, Ba, Mons, Brozo - we're a wacky bunch, but thanks for letting me be my gopher-self. Love you.

To my San Jose, Amherst, New York, Melbourne, Michigan and SoCal crews: I love you all. How I wish we lived closer to each other, but I'm so grateful you are all still a part of my life and thank you so much for supporting me, influencing me, and loving me. I'd like to especially thank a few of my oldest friends for their immense support: Kelly, Karen, Jenny and Kevin. Kelly and Karen, I've known you both for 20+ years now; at this point, you've been in my life longer than you've been out. Thanking you for your love and support doesn't seem enough, but it's a start. Thank you for always providing me with a place to crash when I needed it, and I'm so thankful we have each other. Jenny, I've known you for over a decade now and you've been one of my biggest cheerleaders since day one. Thank you for always trying to make my life easier, looking out for me, and doing virtual work sessions with me. And to Kevin, we've now known each other for a decade but it seems like yesterday when we were wide-eyed kids ready to explore the city. Thank you for always providing me with a safe space and being ready for an adventure. Love you all.

And finally, thank you to Matt: my number one IT guy, my favorite crazy cat man and my partner over the past five years. Thank you for your support, for going through this journey with me despite its ups and downs. Thank you for your patience, for tolerating my alarms, for listening to me figure out ”science” in the middle of the night despite how brain-dead you were. Thank you for always trying to engineer a solution to my problems; thank you for all the sunset chases. I can't imagine doing this PhD without you because you were such an integral part of my experience. You've been my biggest supporter during this journey – thank you from the bottom of my heart and I love you!

VITA

Michelle Nguyen Ngo

EDUCATION

Doctor of Philosophy in Mathematical, Computational and Systems Biology **2023**
University of California, Irvine *Irvine, CA*

Master of Science in Biostatistics **2017**
University of Michigan, Ann Arbor *Ann Arbor, MI*

Bachelor of Arts in Mathematics, Statistics **2015**
Amherst College *Amherst, MA*

RESEARCH EXPERIENCE

Graduate Research Assistant **2018–2022**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant, Statistics **2022**
University of California, Irvine *Irvine, California*

Teaching Assistant, Mathematics **2019**
University of California, Irvine *Irvine, California*

ABSTRACT OF THE DISSERTATION

Leveraging Latent Representations to Investigate Biological Processes

By

Michelle Nguyen Ngo

Doctor of Philosophy in Mathematical, Computational and Systems Biology

University of California, Irvine, 2023

Professor Babak Shahbaba, Chair

The advent and rapid adoption of single cell RNA sequencing technology have ushered in an era of biological breakthroughs at the cellular level. Despite the biological and technical variation in addition to computational challenges, the ability to isolate individual cells and generate their sequencing libraries have enabled researchers to investigate topics such as the discovery of cell sub-populations and ability to infer transcriptional dynamics. On this note, we will explore the latent representations of the single cell RNA sequencing data expression matrix to investigate two complex processes: clonal hematopoiesis and circadian rhythms.

For clonal hematopoiesis, we examine the inflammatory response of a patient with myeloproliferative neoplasms (MPN). Here, we present a computational approach that first identifies ranked groups of differentially expressed genes and then uses those groups to cluster the hematopoietic stem and progenitor cells (HSPCs). We confirm that the MPN patient shows more response to inflammation than the unaffected patient. For circadian rhythms, we present a new framework that accurately predicts the circadian time of transcriptomics samples. We show that this framework can aid in the development of circadian precision medicine disease management plans and that it can also provide evidence of circadian heterogeneity in different cell types.

Chapter 1

Introduction

1.1 Genomics and latent representation

Rapid development of next-generation (or massively parallel) sequencing technology provides an opportunity to characterize individual cells and their gene expression patterns rather than bulk populations, where the gene expression patterns are an average across all the cells. In 2009, Tang *et al.* [113] adapted the complementary DNA (cDNA) amplification technique for compatibility with high-throughput sequencing technology to publish the first single-cell RNA (scRNA) sequencing study [111] and since then, many new sequencing protocols (e.g., Smart-seq2 [87], Drop-seq [73] and CEL-seq2 [49]) and commercial platforms (e.g., Fluidigm C1, Clontech iCell8 and 10x Genomics Chromium) have emerged.

Several papers have reviewed and compared the foremost protocols [141, 111, 138]. The first step of most single-cell sequencing protocols is the isolation of cells from a tissue sample, which then determines the throughput of the method [16]. These isolated single cells are lysed to maximize the capture of RNA molecules and then the following procedures are generally executed: reverse transcription, synthesis of the second cDNA strand, cDNA amplification, preparation of the cDNA sequencing library and pooling of the cDNA sequencing libraries [47].

Due to the samples each being an individual cell and technical challenges during the sequencing process, the resulting scRNA sequencing data expression matrix, which represents the number of transcripts observed for each gene in a single cell, is extremely noisy and potentially inaccurate. There are three main sources of variation in scRNA sequencing: technical, allele-intrinsic and allele-extrinsic variation [122]. Technical variation arises from differences in cell integrity and execution of single-cell sequencing protocols, such as the conversion to cDNA and synthesis of the second cDNA strand. Allele-intrinsic variation encompasses the inherent stochasticity of the molecular mechanisms controlling gene expression, leading two identical copies of a given gene in the same intracellular environment to have uncorrelated expressions [32]. Allele-extrinsic variation emerges from sources extrinsic to the process of transcription since gene expression is also controlled by the presence and activity of regulators and other molecules [122, 32]. One predominant feature of scRNA sequencing data arising from these variations is the large amount of zeros attributed to a phenomenon known as “dropouts”. This occurs when genes that are highly expressed in one cell are instead read as unexpressed (zero expression) during the sequencing process and has led to widespread inconsistencies about the source and interpretation of zeros in scRNA sequencing data [99]. Many methods have been developed to address the alleviation of these technical variations and a selection are reviewed in [122, 5].

Despite these challenges, the ability to isolate single cells and generate their sequencing libraries have enabled novel cell-specific biological breakthroughs such as the discovery of sub-populations or rare types of cells, the tracking of cell development trajectories, the identification of tumor development and heterogeneity, and the ability to infer transcriptional dynamics and regulatory networks [138]. To this end, our lab is interested in utilizing the latent representation of the scRNA sequencing data expression matrix to investigate two complex biological processes: clonal hematopoiesis and circadian rhythms.

1.2 Hematopoiesis application

The first biological process of interest, hematopoiesis, is the lifelong process of formation and maintenance of blood cells and has two developmental waves: primitive and definitive. In mammals, primitive hematopoiesis begins in the blood islands in the embryonic yolk sac and is characterized by the production of large nucleated erythrocytes to facilitate tissue oxygenation [59, 57]. However, the primitive wave is temporary and switches to definitive hematopoiesis at different time points of development depending on the species. For example, in mice, primitive hematopoiesis begins on embryonic day 7 and switches to definitive hematopoiesis between embryonic days 10 and 11 [59]. Definitive hematopoiesis is characterized by the presence of enucleated erythrocytes and generates hematopoietic stem cells (HSCs) and multipotent progenitors (MPPs). Unlike the erythroid progenitors produced during primitive hematopoeisis, these stem and progenitor cells are multipotent and have the capability to renew [57]. In addition to the functionally distinct progenitors, the principal location of definitive hematopoiesis also shifts from the yolk sac briefly to the fetal liver before migrating to the bone marrow. As we focus only on definitive hematopoiesis, any reference to hematopoiesis henceforth refers definitive hematopoiesis not primitive hematopoiesis.

The hematopoietic system is extensively organized into a cellular hierarchy, with HSCs at the top. During hematopoiesis, HSCs can divide either symmetrically to produce two daughter HSCs or asymmetrically to produce an HSC and daughter cell primed for differentiation into functional, mature blood cells such as lymphocytes and granulocytes [9]. In adult humans, approximately one million mature blood cells are produced per second [102]. Thus, maintaining the balance of hematopoiesis is a critical and highly controlled process depending on the individual's needs. Under homeostatic conditions, an adequate supply of HSCs must be maintained for the lifespan of the individual as mature blood cells are predominantly short-lived and need to be continually replenished, and in the event of hematopoietic stress such as an infection, the demand for increased production of mature blood cells needs to be met.

However, over time, dividing HSCs may acquire mutations that are passed onto the next generation of cells. While most somatic mutations do not have an impact on the function of stem cells, a phenomenon called clonal hematopoiesis can occur where the mutation enables the HSC clone to have a selective advantage over other clonal lineages and eventually have a detectable presence in the population [9].

One such myeloid malignancy arising from clonal hematopoiesis is myeloproliferative neoplasms (MPNs). MPNs are a group of rare chronic cancers in the bone marrow; there are three major MPN subtypes: myelofibrosis (MF), polycythemia vera (PV) and essential thrombocythemia (ET). Each of these subtypes present with different clinical characteristics, symptoms and prognoses. Despite the clinical distinctions, all three subtypes have been shown to activate mutations in the *JAK2*, *MPL*, and *CALR* genes as well as the Janus kinase (JAK)-signal transducer and activator of transcription (STAT) signaling pathway [43]. For patients with the PV subtype, approximately 95% of them have detectable $JAK2^{V617F}$ mutations and the rest have *JAK2* exon 12 mutations. The $JAK2^{V617F}$ mutation is also detectable in approximately 50 - 60% of patients with the ET or MF subtypes [43, 53]. Thus, one current therapeutic approach is to use JAK inhibitors like Ruxolitinib [119], which inhibit the JAK1 and JAK2 kinases.

Numerous studies have shown that chronic inflammation and most types of cancer are linked, as chronic inflammation can create a microenvironment rich in inflammatory cells and growth or survival factors that are favorable to the development of cancer [19, 92, 106, 140]. For MPN in particular, inflammation plays a key role in its development, progression and symptomatic burden. MPN patients typically have increased pro-inflammatory cytokines such as tumor necrosis factor alpha (TNF-α) and interleukin 6 (IL-6) [77]. Futhermore, hematopoietic progenitors with the *JAK2*^{V617F} mutation are resistant to inflammation [36], and many experiments show that activation of the JAK-STAT pathway promotes MPN progression [18, 65, 76, 53].

Thus, our lab is interested in utilizing statistical methods to leverage the information obtained from current sequencing technologies to determine how inflammation affects MPN patients compared to normal patients.

1.3 Circadian application

The second biological process of interest, circadian rhythms, is an endogeneous process governed by a biological clock with a period of roughly 24 hours. In mammals, the circadian clock is primarily established in a hierarchical order: the suprachiasmatic nucleus (SCN) in the brain acts as the central clock ("pacemaker"), and synchronizes and sets the circadian rhythm in the peripheral tissues through neuronal and hormonal signals. These clocks are a biochemical oscillator powered by transcription-translation loops. The CLOCK:BMAL1 transcription complex activates transcription of *Period* (*Per1*, *Per2*) and *Cryptochrome* (*Cry1*, *Cry2*) genes. At high levels, PER and CRY then form a repressor complex to inhibit their own expression by repressing CLOCK:BMAL1 activity. This negative feedback loop, along with other post-translational modifications, generates a very robust 24 hour oscillation of clock protein levels and activity.

Although the molecular mechanisms of this rhythm are cell autonomous and highly conserved in the SCN and peripheral cells, they may respond and adjust to external cues such as light [12] or feeding [86, 124]. These inputs can then lead to discoordination between internal and external circadian patterns, which can greatly affect the general well-being of organisms. One commonly known circadian disruption is jet lag, which is a temporary disorder that occurs when an individual crosses one or more time zones. The rapid change in the sleep-wake cycle causes an individual's internal clock to be out of sync with cues such as light in the new time zone [121]. Other health issues often correlated with circadian disruption are mental health problems such as depression and anxiety [123], the severity of metabolic diseases such as diabetes and obesity [61, 89, 120], and prognosis of cancers such as melanoma and breast [75, 104, 44, 134, 101, 45]. As circadian rhythm is so intertwined with an individual's health, circadian precision medicine aims to consider a patient's circadian rhythm when developing preventative care or disease management plans. For example, short half-life statins for reducing cholesterol work best when taken before bedtime [137]. Thus, the first and most important step in circadian precision medicine is to determine the internal circadian time of the patient or tissue of focus. To this end, our lab is interested in developing efficient, robust pipelines to predict circadian time of transcriptomic data.

1.4 Thesis Organization

In Chapter 2, we introduce an experiment to investigate the transcriptomic response to inflammation in an MPN patient compared to a normal patient. Using a Bayesian mixture model with shrinkage priors, we identify differentially expressed genes between the original samples and the samples treated with a pro-inflammatory factor for each patient. Then in Chapter 3, we present a computational pipeline using two Dirichlet process mixture models to simultaneously cluster the rows and columns of a matrix into homogeneous submatrices. We apply this to single cell RNA sequencing data to determine if we can identify groups of genes that are highly expressed for an inferred group of cells. Due to computational challenges, in Chapter 4, we use the gene groups identified in Chapter 2 and extend the analysis by leveraging the expression profiles obtained via single cell sequencing. Here, we present a pipeline to cluster the sequenced cells and then examine the differential expression between the original samples and treated samples for each patient given the cell clusters. Finally, in Chapter 5, we shift our focus to circadian rhythms and present a pipeline to predict the circadian time of transcriptomic data.

Chapter 2

Investigating the effect of TNF-α on HSCs in an MPN patient

2.1 MPN Background

Myeloproliferative neoplams (MPN) are a collection of rare blood cancers resulting from mutations in hematopoietic stem and progenitor cells (HSPCs). The World Health Organization (WHO) dichotomizes MPNs into BCR-ABL1-positive or BCR_ABL1-negative depending on the presence or absence of this gene [37]; of the BCR-ABL1-negative MPNs, there are three subtypes that make up what is referred to as the "classical MPNs": essential thrombocythemia (ET), myelofibrosis (MF) and polycythemia vera (PV). Each of the three subtypes has differing clinical characteristics and molecular features, while sharing some similarities in pathogenesis and symptoms – making diagnosis challenging. In general, MPN causes an overproduction of blood cells in the bone marrow, leading to increased spleen size, fatigue and bone marrow failure. Elevated levels of pro-inflammatory cytokines, particularly tumor necrosis factor alpha (TNF-α), worsen MPN progression, severity and symptom burden. The standard therapy to treat MF patients (and later PV patients) have been JAK2 inhibitors, but this therapeutic has been shown to alleviate the symptoms rather than treat the disease. Similarly, treatments for ET and PV are aimed at reducing the risk of thromboembolic and cardiovascular complications and symptom burden. Currently, the only cure for MPN is a bone marrow transplant, but due to the high mortality risk, this procedure is usually reserved for younger patients with a specific subtype and still achieves a poor outcome [43].

Despite the clinical distinctions, all three subtypes have shown to activate mutations in the *JAK2*, *MPL* and *CALR* genes and JAK-STAT signaling pathway. Notably, the $JAK2^{V617F}$ mutation is found in approximately 95% of patients with the PV subtype, and 50-60% of patients with the ET or MF subtypes [43, 53]. The $JAK2^{V617F}$ mutation increases resistance to inflammation in progenitors and these mutant cells also produce pro-inflammatory cytokines such as TNF-α while inducing surrounding normal cells to produce inflammatory cytokines [77]. This presents an issue for MPN patients, as MPN is also suggested to be a chronic inflammatory disease in addition to neoplastic disorder, and patients usually present with very high levels of inflammation [50, 126, 3, 35].

One of the pro-inflammatory cytokines found in high levels in MPN patients is TNF-α, which upregulates multiple pro-inflammatory proteins through activating the NF-κB (nuclear factor kappa-light-chain-enhancer of activated B cells) and MAPK (Mitogen-Activated Protein Kinase) pathways [126]. TNF-α plasma concentrations were found to be 10-, 5-, and 4-fold higher in MF, PV and ET patients than control groups respectively [36], and at least three more studies confirm that TNF-α is elevated in MPN patients and that the pro-inflammatory cytokine may play a role in MPN development [126]. As further investigation is needed to determine the exact role and mechanism of TNF-α in MPN, our lab would like to examine the differential gene expression of MPN HSCs in response to TNF-α induced inflammation compared to a healthy (unaffected) control.

2.2 Experiment

We used droplet-based single cell RNA sequencing (scRNA seq) to investigate transcriptional profiling in primary human bone marrow hematopoietic stem and progenitor cells. First, we purified mononuclear cells from fresh bone marrow aspirates from one female MPN patient (Polycythemia Vera with 71% $JAK2^{V617F}$ allele burden) as well as one unaffected, age-matched, male individual then sorted Lineage-/CD34+/CD38- hematopoietic progenitors by flow cytometry. Immediately following sorting, half of the cells were stimulated with 50ng/ml tumor necrosis factor alpha (TNF-α) for 4 hours at 37°C while the other half of cells were used as unstimulated controls. We then utilized the 10X Chromium platform to generate single-cell droplets for the 8,129 total cells from the unaffected individual and 33,299 total cells from the MPN patient. Figure 2.1 outlines the experiment described above.

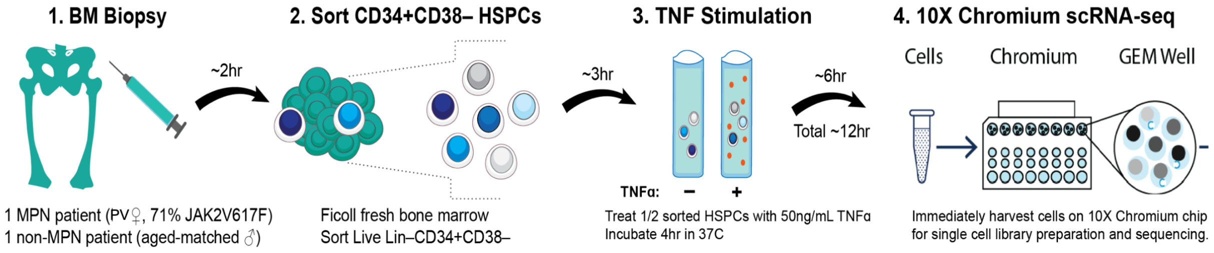

1. BM Biopsy
~2hr
2. Sort CD34+CD38– HSPCs
~3hr
3. TNF Stimulation
~6hr
Total ~12hr
4. 10X Chromium scRNA-seq
Cells
Chromium
GEM Well
TNFα: – +
1 MPN patient (PV♀, 71% JAK2V617F)
1 non-MPN patient (aged-matched ♂)
Ficoll fresh bone marrow
Sort Live Lin–CD34+CD38–
Treat 1/2 sorted HSPCs with 50ng/mL TNFα
Incubate 4hr in 37C
Immediately harvest cells on 10X Chromium chip
for single cell library preparation and sequencing.

Figure by Brianna Hoover, Fleischman Lab, UCI

Figure 2.1: Experimental schematic. We took fresh bone marrow aspirates from one female MPN patient and one age-matched, unaffected male individual and sorted hematopoietic progenitors by flow cytometery. We divided each sample in half and stimulated one half of the cells with 50ng/ml tumor necrosis factor alpha for 4 hours at 37°C. After spinning the cells and washing them, we immediately harvest the cells for synthesis.

2.3 Differential Gene Analysis Background

To determine the effect of a pro-inflammatory cytokine on MPN, in particular TNF-α, we need to determine which genes are expressed at different levels between the unstimulated ("untreated") and stimulated ("treated") samples in both patients. Although our experiment yielded scRNA seq data, we are initially interested in the effect of TNF-α on the HSPCs overall. In Chapter 4, we will examine the effect of TNF-α after accounting for the heterogeneity of the samples. Thus, we can then formulate our problem into two separate two group tests, where we independently detect the differentially expressed genes in each patient between the two samples.

Many tools have been developed for differential gene expression analysis on bulk RNA sequencing data. Several of these, such as DESeq, DESeq2 [70] and edgeR [95], in addition to classical statistical methods such as the Wilcoxon and t-test, can also be applied to scRNA seq data. However, scRNA seq differs from bulk RNA seq in several ways with the most crucial difference being that bulk RNA seq measures the average gene expression across the cell populations in a sample while scRNA seq measures the gene expression of each individual cell in a sample. The number of bulk RNA seq samples (typically the number of replicates per condition) is also much smaller than the number of individual cells sequenced in scRNA seq. To address these differences, several methods such as waddR [100] and MarcoPolo [63] have been developed for differential gene expression analysis on scRNA seq data specifically. While there is no general consensus on which method is the "best" for determining differential genes with scRNA seq data [125, 83], all these methods find biologically meaningful differential genes and many have overlapping results. However, they all determine whether a gene is differential or not by a cutoff.

We are interested in not only finding the differential genes but also in ranking the groups of differentially expressed genes. To do so, we begin with a two group test to compare the genes between the unstimulated and stimulated samples in each patient. In 2008, Efron incorporated Bayesian methods to the classical two group model to control the false discovery rate (FDR) in large-scale testing situations [31] such as microarrays. Extending this idea, Denti *et al.* [23] proposed a Bayesian mixture model with shrinkage priors to partition the region of interest into varying degrees of relevance; in our case, we will be partitioning the genes into varying degrees of differentiation.

2.4 Method

2.4.1 Pre-processing

For quality control, we filter out cells with mitochrondrial expression greater than 7% and any suspected dead or empty cells, doublets or multiplets. Each retained cell has at least 10 genes expressed, and we remove any genes where there are no (zero) expression in both the untreated and treated samples. Genes with zero expression in one sample but expressed in the other sample are kept for further analysis. Each data set is then normalized to account for any bias such as differences in sequencing depth per cell. The gene counts for each cell are divided by the total counts for that cell and then multiplied by a scale factor of 10^4. The normalized count data X is then log-transformed: $\log_2(X + 1)$.

2.4.2 Calculating the distance between each sample

To compare the unstimulated sample with the stimulated sample for each patient, we use the two sample Wilcoxon rank sum test. This is a non-parametric alternative to the two sample t-test and checks if the distributions of the two groups differ significantly from each other. That is, suppose the distribution of expression for a given gene in the unstimulated sample is A and the distribution of expression for the same gene in the stimulated sample is B. The Wilcoxon test attempts to detect whether B departs from A. For each gene, the two-sided Wilcoxon rank sum test yields a W statistic, p-value and z-score. However, for modeling convenience, we will work with the z-scores instead of the p-values. The p-values are transformed into z-scores using the percentage point function outlined in [129]. For each sample, the transformed z-scores are then centered to have a mean of zero. In the following sections, we outline how we will group the transformed z-scores to derive a ranking of the groups of differentially expressed genes.

2.4.3 Dirichlet Process

The Dirichlet process (DP) is a stochastic process that defines a distribution over distributions such that its marginal distributions are Dirichlet distributed [34, 103]. Formally, let G_0 be a distribution over some probability space Ω and α be a positive real number. Then a stochastic process G is said to be a Dirichlet process with parameter α if for any finite partition $(A_1, ..., A_k)$ of the probability space Ω, the vector $(G(A_1), ..., G(A_k))$ is random and has a Dirichlet distribution with base distribution G_0 and concentration parameter α. That is, $G \sim DP(\alpha, G_0)$ if $(G(A_1), ..., G(A_k)) \sim Dirichlet(\alpha G_0(A_1), ..., \alpha G_0(A_k))$. The concentration parameter α can then be thought of as a precision (inverse variance) parameter, and the base distribution G_0 can be thought of as a mean parameter (i.e., the mean of the DP).

Given its partitioning properties, the Dirichlet process is often used in Bayesian non-parametric approaches such as Dirichlet process mixture models. Dirichlet process mixture models (DP-MMs) remove the need to pre-specify the number of clusters by placing a DP prior over the cluster parameters and can be intuitively thought of as an infinite dimensional generalization of finite mixture models that incorporates automatic model selection. Consider a Bayesian mixture model with K clusters and then extending $K \to \infty$:

$$\begin{aligned} x_i|z_i, \theta_k &\sim F(\theta_{z_i}) \\ z_i|\pi &\sim Multinomial(\pi) \\ \theta_k &\sim G_0 \\ \pi|\alpha &\sim Dirichlet\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right) \end{aligned}$$

Here $x_1, \ldots, x_n$ is the observed data and drawn from a mixture of distributions with the form $F(\theta)$, θ is the mixing distribution over G, and z is the cluster assignments for each observation [84]. However, taking the limit of the model such that $K \to \infty$ is not trivial; instead, one can use other constructions to emulate the Dirichlet process. One construction of the Dirichlet process uses the stick breaking representation [103] to define the cluster weights. For the stick breaking process, suppose we have a stick of length one. We can break off a piece of the stick, with probability $z_1 \sim Beta(1, \alpha)$, so that $\pi_1 = \beta_1$. We continue breaking off pieces from the remainder of the stick according to $z_k \sim Beta(1, \alpha)$, and π_k is calculated from the broken off piece:

$$\begin{aligned} z_k &\sim Beta(1, \alpha) \\ \pi_k &= z_k \prod_{j=1}^{k-1} (1 - z_j) \end{aligned}$$

By construction, if the number of clusters $K \to \infty$, we have $\sum_{k=1}^{\infty} \pi_k = 1$. The stick-breaking process orders mixture components such that the weights are stochastically decreasing for each index k [38]. We choose a high index N as an upper bound on the number of clusters and use a truncation approximation to the Dirichlet process. Note that the stick breaking process is also referred to as the Griffiths-Engen-McCloskey (GEM) distribution.

The stick breaking representation of the Dirichlet process mixture model is as follows:

$$\begin{aligned} x_i|z_i, \theta_k &\sim F(\theta_{z_i}) \\ z_i|\pi &\sim Multinomial(\pi) \\ \theta_k|G_0 &\sim G_0 \\ \pi|\alpha &\sim SB(\alpha) \end{aligned}$$

2.4.4 Horseshoe Mixture Model

Using the model presented in [23], we aim to cluster the centered z-scores into multiple groups of varying significance. The group containing z-scores with the lowest variance will represent the null distribution and thus can be interpreted as the scores corresponding to unchanged genes. The other groups will represent varying degrees of differential expression and can be ranked accordingly. At minimum, we will obtain two groups: (1) a group of non-significant z-scores corresponding to unchanged genes and (2) a group of significant z-scores corresponding to the differentially expressed genes in response to TNF-α induced inflammation.

From [23], we consider the following model for the vector of n centered z-scores $\{z_i\}_{i=1}^n$ for each gene i:

$$\begin{aligned} z_i &= I_n \beta_i + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

where $\beta = \{\beta_i\}_{i=1}^n$ is the vector of means, $\epsilon = \{\epsilon_i\}_{i=1}^n \sim \mathcal{N}_n(0, \Sigma)$ is the noise term, and $\Sigma = \sigma^2 I_n$ for simplicity. To impose regularization on the means (i.e., shrink the means corresponding to the unchanged genes to zero), we assume a discrete mixture of continuous scale mixtures of Gaussians as the prior distribution for β:

$$\begin{aligned}
z_i|\beta_i, \sigma^2 &\sim \mathcal{N}(\beta_i, \sigma^2) \\
\beta_i|\tau, \lambda, \zeta_i, \sigma^2 &\sim \mathcal{N}(0, \sigma^2 \cdot \tau^2 \cdot \lambda_{\zeta_i}^2) \\
\zeta_i|\pi &\sim \sum_{l=1}^{L} \pi_l \delta_l(\cdot) \\
\lambda &\sim \mathcal{C}^+(0,1) \\
\pi &\sim SB(\alpha) \\
\sigma^2 &\sim \Gamma(1,1) \\
\tau &\sim \text{fixed in fitting} = 0.0001
\end{aligned}$$

Here, π is the vector of mixture weights, $\tau \in R^+$ is the global shrinkage parameter, $\lambda = \{\lambda_{\zeta_i}\} \in R^+$ is the local (mixture component) shrinkage parameters. The augmentation of the model with latent membership labels $\{\zeta_i\}_{i=1}^n$, where $\zeta_i \in \{1, ..., L\}$, links each z-score with a cluster and enables us identify and rank the different groups of z-scores by significance.

To fit this model, we use a Bayesian non-parameteric approach; we adopt the stick breaking representation of the Dirichlet process as the prior distribution for the mixture weights π and fix the value of τ^2 to 0.0001. We ran 10,000 iterations for the burn-in period and used the next 10,000 iterations for inference. More details on the model, posterior inference, and post-processing can be found in [23].

2.4.5 Post-processing

Once we have the posterior samples, we can partition the z-scores into groups of similar significance. To do so, we take the cluster membership labels $\zeta^{(t)} = \{\zeta_1^{(t)}, ..., \zeta_n^{(t)}\}$ for iteration $t = 1, ..., T$ and construct the posterior similarity matrix (PSM). Each entry $[i, j]$ in the PSM contains the posterior probability that z_i for gene i and z_j for gene j are clustered together:

$$\text{PSM}_{i,j} = \frac{\sum_t^T 1_{(\zeta_i^{(t)} = \zeta_j^{(t)})}}{T}, \text{ for } i, j = 1, ..., n$$

Since the PSM yields estimates of the proportion of times two z-scores are clustered together during the T MCMC iterations, we subtract the PSM from 1 to obtain a dissimilarity (or distance) matrix: $1-$ PSM. Hierarchical clustering using Ward linkage is then applied to the dissimilarity matrix to partition the z-scores. The number of partitions chosen is the mode of the total number of clusters identified over all the MCMC iterations.

We define a cluster as a significant cluster if the minimum centered z-scores in that cluster have a magnitude greater than two. All other clusters are deemed as “non-significant” or “non-differential”.

2.5 Results

2.5.1 Gene Groups for the Unaffected Patient

From the framework described in Chapter 2, we obtain 16 groups (clusters) of genes for the unaffected patient. Table 2.1 below shows how many genes are assigned to each cluster for both the unaffected patient and MPN patient. For both patients, the non-significant clusters (those with smaller numbered labels) contain the majority of the genes which corroborates previous findings that not all genes will contain information we can leverage to determine whether our findings have any biological significance.

Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**
Patient 013 (unaffected)	1050	2979	6525	856	585	913	1141	692	402	204	359	189	251	315	559	75	—
Patient 025 (MPN)	6113	3197	1474	900	1037	576	856	1282	1257	789	399	354	134	155	1140	238	130

Table 2.1: The number of genes in each cluster for the unaffected and MPN patients.

Figure 2.2 shows the distribution of centered z-scores prior to clustering, a plot of the β's (means) against the centered z-scores, and a plot of the centered z-scores colored by its assigned gene group. Note that the majority of the centered z-scores are around zero, indicating there is no statistical difference in the expressions of those genes between the stimulated and unstimulated groups for this patient. When we inspect Figure 2.2, we see that the framework has imposed regularization on the means and shrunk the genes that did not change much in expression to zero. We can also visually examine the gene cluster assignments in Figures 2.2D. The cluster labels are ranked in order of significance; the lower the label, the more the centered z-scores are close to zero. In the "non-significant" clusters, the peak of the distributions are centered roughly around zero while the "significant" clusters are more bimodal, with no centered z-scores around zero. Since we define significant clusters as any cluster where the minimum centered z-scores have a magnitude greater than 2, we obtain nine significant gene clusters for the unaffected patient. We will examine the genes in the most significant cluster for both the unaffected and MPN patients in Section 2.5.3.

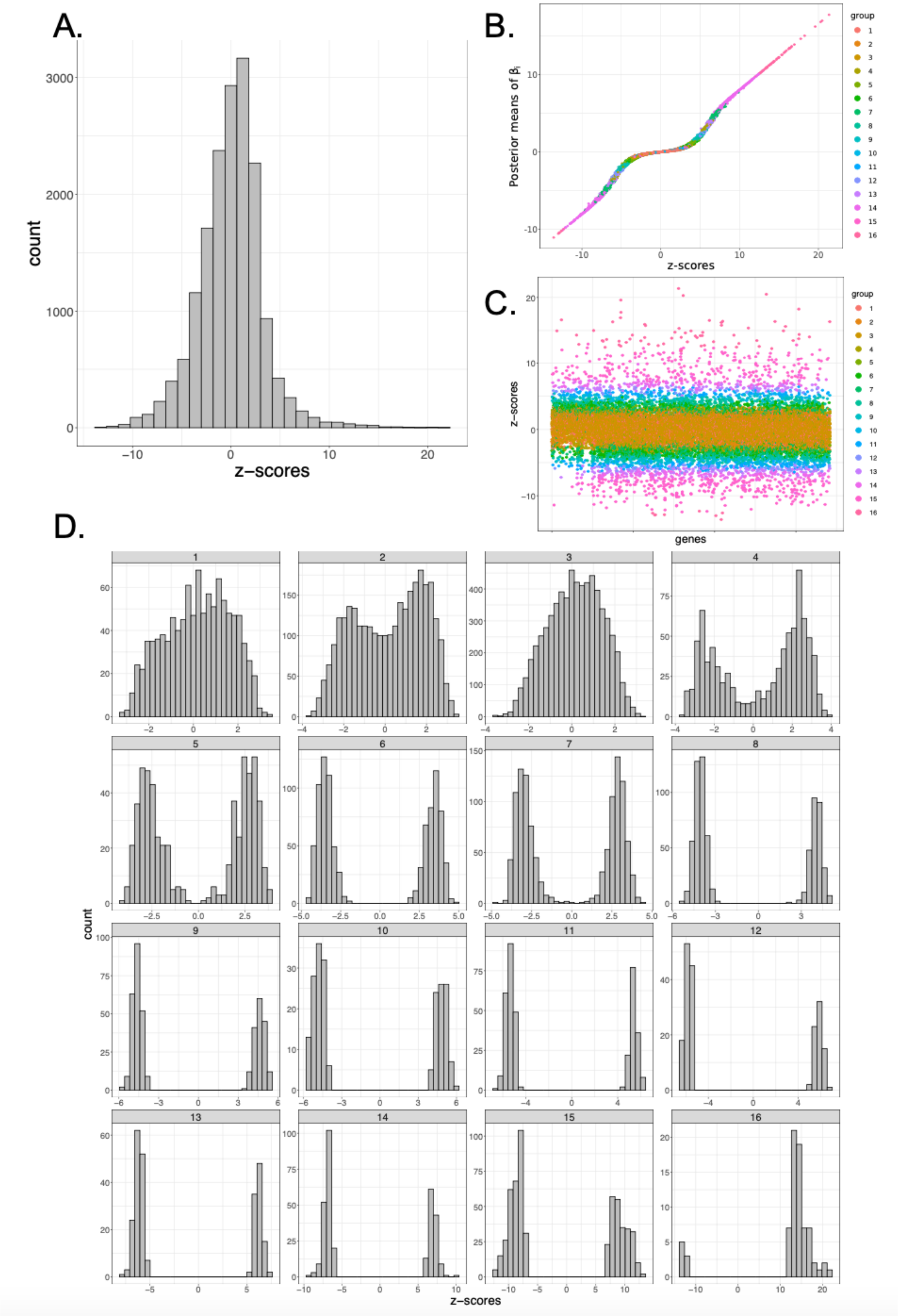


Figure 2.2: A. Distribution of centered z-scores prior to clustering; B. a plot of the posterior means of the means against the centered z-scores; C. a plot of the centered z-scores colored by its assigned gene group; D. distributions of the centered z-scores for each gene cluster.

2.5.2 Gene Groups for MPN patient

Similarly, we obtain 17 groups (clusters) of genes for the MPN patient. Figure 2.3 shows the distribution of centered z-scores prior to clustering, a plot of the β's (means) against the centered z-scores, and a plot of the centered z-scores colored by its assigned gene group. Again, note that the majority of the centered z-scores are around zero, indicating there is no statistical difference in the expressions of those genes between the stimulated and unstimulated groups for this patient. When we inspect Figure 2.3, we see that the framework has imposed regularization on the means and shrunk the genes that did not change much in expression to zero. We can also visually examine the gene cluster assignments in Figure 2.3D. The cluster labels are ranked in order of significance; the lower the label, the more the centered z-scores are close to zero. In the "non-significant" clusters, the peak of the distributions are centered roughly around zero while the "significant" clusters are more bimodal, with no centered z-scores around zero. Since we define significant clusters as any cluster where the minimum centered z-scores have a magnitude greater than 2, we obtain fourteen significant gene clusters for the MPN patient. Compared to the unaffected patient, the difference in gene expression between the unstimulated and stimulated samples are much greater in the MPN patient. The centered z-scores for the unaffected patient range from approximately -10 to 20 while the range for the MPN patient is an order of magnitude larger: approximately -150 to 100. We will examine the genes in the most significant cluster for both the unaffected and MPN patients in next section.

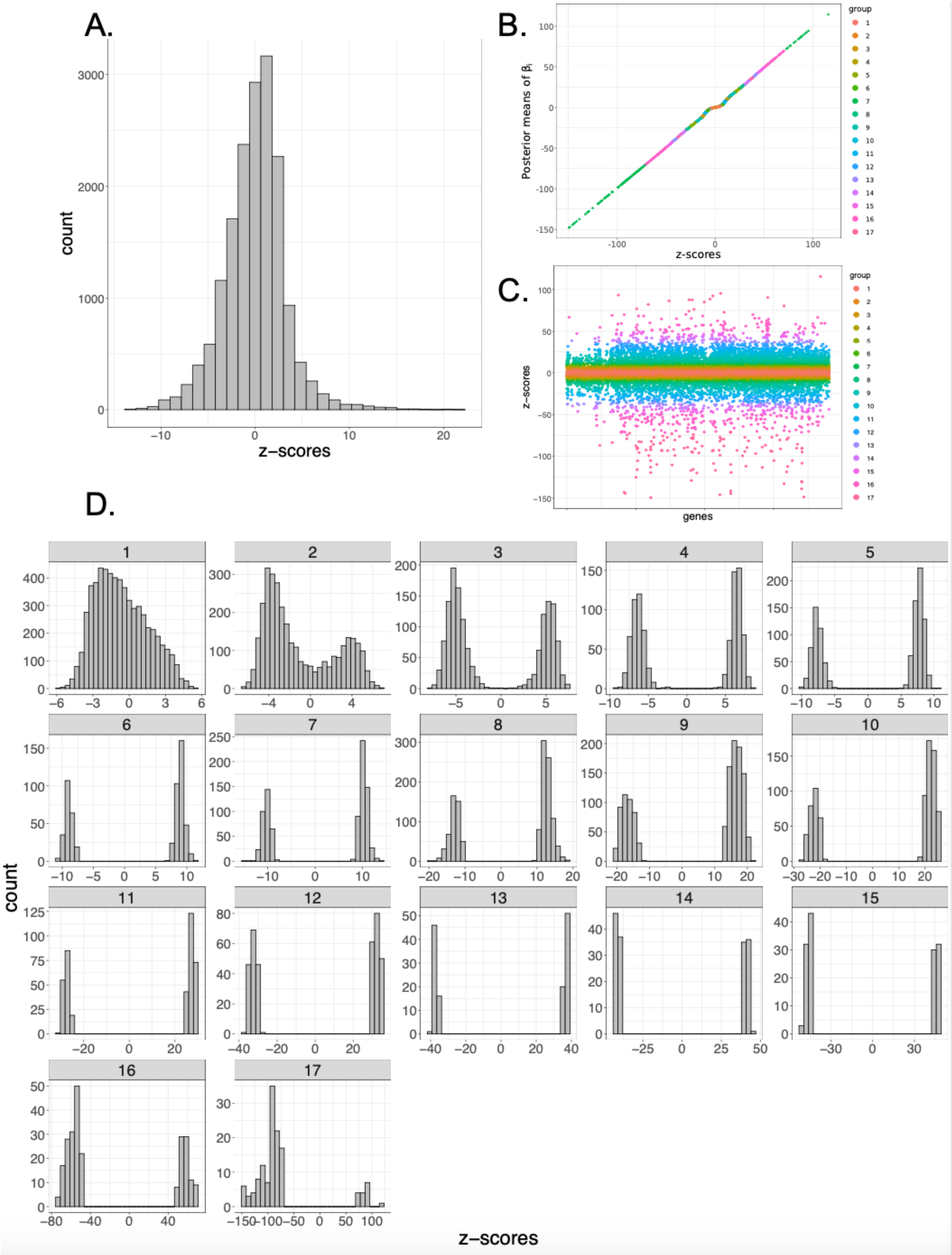


Figure 2.3: A. Distribution of centered z-scores prior to clustering; B. a plot of the posterior means of the means against the centered z-scores; C. a plot of the centered z-scores colored by its assigned gene group; D. distributions of the centered z-scores for each gene cluster.

2.5.3 Examining the Gene Clusters' Biological Relevance

In the previous two sections, we described the results of the framework and how many significant gene clusters we found in the unaffected and MPN patients. Here, we examine the most significant clusters (i.e., the ones with the highest cluster labels) to determine if the clusterings have any biological relevance and/or significance.

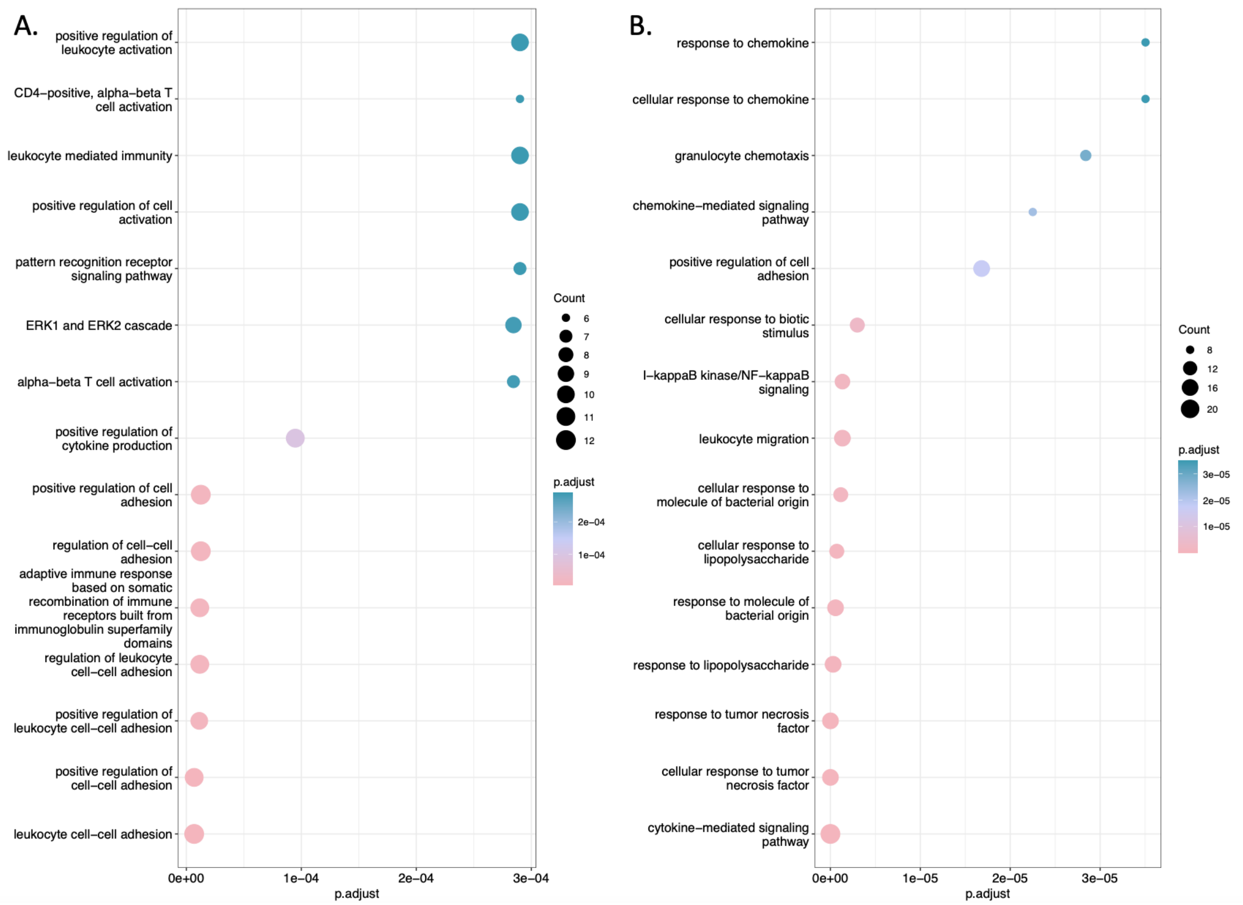


Figure 2.4: Top 15 biological processes for the most significant cluster in the unaffected patient (A) and MPN patient (B), sorted by their adjusted *p*-values.

Figure 2.4 shows the top 15 biological processes for the most significant gene cluster in each patient. The most important cluster in the unaffected patient is cluster 16, where the top 10 differentially expressed genes (with respect to the magnitude of their log2 fold change) are: *PLAU, IL2RA, ICAM1, MIR3142HG, NFKB2, MAFF, KDM6B, HIVEP2, CD83* and

IER5. These genes are enriched for *positive regulation of transcription by RNA polymerase II* (GO:0045944), as well as the NF-κB signaling pathway (KEGG 2021 Human) and TNF-alpha effects on cytokine activity, cell mobility, and apoptosis (BioPlanet 2019).

In the MPN patient, the most significant cluster is 17 and the top 10 differentially expressed genes here (again, with respect with to the magnitude of their log2 fold change) are: *CXCL8, NFKBIA, MIR155HG, IER3, BIRC3, TNFSF10, TNFAIP8, SOD2, NFKB1* and *CXCL3*. These genes are enriched for *cytokine-mediated signaling pathway* (GO:0019221) and *negative regulation of apoptotic process* (GO:0043066). They also significantly correspond to several expected pathways: the lipid and atherosclerosis and NF-κB signaling pathways (KEGG 2021 Human); TNF-alpha effects on cytokine activity, cell mobility, and apoptosis (BioPlanet 2019); and IL-18 signaling pathway WP4754 (WikiPathway 2021 Human).

Our framework confirms that the MPN patient has a higher response to inflammation than the unaffected patient [35]. Craver *et al.* performed a preliminary analysis on the same data and suggested a MPN cells have a dampened apoptotic response to TNF-α [21]; we can corroborate this statement. Firstly, the term *negative regulation of apoptotic process* (GO:0043066) was enriched for the MPN samples. Furthermore, the unaffected patient's stimulated sample highly expressed genes involved in the caspase cascade (e.g., *CASP2, CASP6, CASP8*) when compared to the unstimulated sample. Our framework assigned the three aforementioned genes to the second-most differential ("significant") cluster; two other caspases were assigned the third and fourth-most differential clusters respectively and the rest were found to not be significantly differentially expressed. In the MPN patient, most of the capases were found to be more highly expressed in the unstimulated sample with the exception of *CASP3* and *CASP7*. However, the highest rank our framework assigned any of the capases in the MPN patient is 14, so the fourth-most differential cluster.

2.6 Discussion

In this computational framework, we investigated the differential gene expression between an untreated sample and a sample treated with TNF-α in an unaffected patient and a MPN patient. We begin by performing a Wilcoxon ranked sum test on each of the genes in the unstimulated and stimulated samples in each patient. After converting the test statistics to z-scores and centering them, we then use a Horseshoe mixture model based approach to simultaneously shrink the means corresponding to the non-differentially expressed genes to zero while clustering all the means. Notably, our framework is able to capture not just the most significant differentially expressed genes but several groups of significant differentially genes and rank these groups. This enables us to explore more differentially expressed genes and how that response differs between the two patients rather than finding just a group of significant differentially expressed genes and taking the top n genes.

In doing so, we were able to confirm that the MPN patient responds more to inflammation than the unaffected patient, as stated in [35]. For the MPN patient, not only does the most significant gene cluster predominantly correspond to inflammatory responses but the difference between the unstimulated and stimulated samples are an order of magnituded larger than the unaffected patient. One next step would be to compare the differential genes between the two patients (i.e., unstimulated samples for the unaffected patient and MPN patient, stimulated samples for the unaffected patient and MPN patient). We expect that the MPN patient will have more inflammatory markers than the unaffected patient in both cases.

Other approaches that determine which genes are differentially expressed given at least two samples can also be applied here. However, if we "ignore" heterogeneity and treat the unstimulated and stimulated samples for each patient where each column (cell) is a "sample", then the dimensionality of these data matrices is a limitation. For example, DESeq2 [70] (and its predecessor DESeq) were developed for bulk RNA seq and have been successfully applied to scRNA seq data. Both are based on the negative binomial distribution and fit a generalized linear model to each column (sample); as we have 3046 genes and 15103 “samples” for each treatment, this method will be computationally intensive. edgeR [95] and limma [94] can also be used for differential analysis and are more efficient for larger samples than DESeq2, but they were not designed to handle more than a couple thousand of samples.

Although our framework is also computationally intensive, we can extract more information such as the degree to which a given gene belongs to a gene cluster. With our large sample size, fitting the Horseshoe mixture model and calculating the posterior similarity matrix are very computationally expensive. As sequencing technologies improve, the ability to scale the inference will be crucial.

2.7 Acknowledgements

This work is a collaboration between the Lowengrub, Shahbaba and Fleischman labs at UC Irvine. Members of this project include: Brianna M. Hoover, Michelle N. Ngo, Quy Nguyen, Gajalakshmi Ramanathan, John Lowengrub, Babak Shahbaba, and Angela G. Fleischman.

Chapter 3

Conjoined Dirichlet Process

Michelle N. Ngo[1], Dustin S. Pluta[1], Alexander Ngo, Babak Shahbaba

[1] Equal contribution

3.1 Foreword

In Chapter 2, we introduced an experiment to investigate how TNF-α affects a normal patient versus a MPN patient. We identified differentially expressed genes between the unstimulated samples (not treated with TNF-α) and the samples treated with TNF-α in each of the two patients. Now we extend that analysis by investigating the differentially expressed genes between the unstimulated and treated samples among the different cell populations in each of the two patients.

Within the HSPCs, there are several different populations with distinct self-renewal and repopulating capacity as well as lineage differentiation. Kleppe *et al.* used single cell profiling to demonstrate that there is cellular heterogeneity in cytokine secretion in the hematopoietic cells of MF patients compared to normal hematopoietic cells [65]. Recently, Tong *et al.* also used single cell profiling to show that *JAK2* mutant HSCs in ET patients are biased toward the megakaryocyte (Mk) lineage, and responsible for expanding the Mk-primed HSC subpopulation in ET patients [116]. Our Lin-/CD34+/CD38- hematopoietic stem and progenitor cells should also display some heterogeneity, at the very least between stem and progenitor cells.

Since single cell data is inherently noisy, a big challenge for researchers is how one should pre-process the expression data to filter out background noise to reflect biological significance since all downstream analyses are dependent on these pre-processing steps. To identify the different cell populations (i.e., cluster the cells), a common and popular approach is to first identify the highly variable genes in the expression matrix of interest [4, 5, 11, 71, 133]. The idea is that only the genes with higher-than-expected variances are informative in clustering the cells into homogeneous populations as they contribute the most to cell-to-cell variability. However, highly variable gene selection may be sensitive to outliers, normalization techniques and depends on the number of selected genes. This may also limit the ability to define rare cell populations as the selected highly variable genes are generally the genes with the highest expression variance across the whole data set of interest which potentially corresponds to the most dominant cell types. For more details and a general comparison on tools for highly variable gene selection, please refer to [133] or for comparison using scRNA seq data from hematopoietic stem and progenitor cells and mature blood cells, see [139].

In this chapter, we explore using information from both the genes and cells simultaneously to create partitions in the data. These partitions can be thought of as joint gene-cell clusters ("biclusters") and help us investigate the heterogeneity in hematopoietic stem and progenitor cells without explicitly selecting for highly variable genes.

3.2 Introduction

Biclustering, or co-clustering, is a technique used for sorting heterogeneous data into homogeneous blocks by allowing for simultaneous clustering of the rows and columns of a matrix. This technique has various important applications, including text mining and biological gene expression analysis. In text mining, biclustering text data from a document corpus allows for identification of document-word combinations with high co-occurrence. Extracted biclusters represent combinations of words and documents that form a (latent) topic. Biclustering has been particularly popular in the past several decades for gene expression microarray analyses. The method is used to group genes into similar conditions to study the functional roles of genes. More recently, biclustering is being used to analyze single cell RNA sequencing data. Here, the method is usually used to study cell proliferation by grouping cells into developmental stages and identifying the genetic drivers for each stage.

Current biclustering methods generally impose restrictive assumptions on the biclustering structure or data-generating mechanisms. However, in real-world applications, which are often exploratory, an appropriate model and bicluster structure can be difficult to specify. To address these limitations in current methods, we propose the Conjoined Dirichlet Process (CDP): a novel, non-parametric probabilistic biclustering method based on dual Dirichlet processes to identify biclusters with strong co-occurrences in both rows and columns. The name of the method derives from its usage of two conjoined Dirichlet process mixture models (DPMMs), akin to conjoined twins (see Figure 3.1). CDP provides the following advantages: 1) the number of biclusters is determined by the data and prior, and does not require selecting a number of clusters *á priori*, 2) fewer modeling assumptions compared to parametric alternatives, 3) estimated biclusters may overlap arbitrarily, and 4) efficient computational methods allow applications to high dimensional data, making applications to text and gene expression data practical.

The paper is organized as follows. In Section 3.3 we describe existing biclustering methods. In Section 3.4 we provide some background on DPMMs, particularly focusing on the parallel MCMC sampler for DPMMs. In Section 3.5 we discuss and provide details of our proposed biclustering method. In Section 5.2 we apply our method to simulated, text, and single cell RNA sequencing data sets, and present the results. Finally, in Section 5.4 we present our conclusion.

3.3 Previous Methods

Briefly, biclustering algorithms are based on four heuristics: greedy, divide-and-conquer, exhaustive enumeration, or distribution parameter identification [85].

Hartigan [48] proposed the first biclustering algorithm in 1972, but the technique was not popular until 2000 when Cheng *et al.* [17] applied it to gene microarray data. Other popular gene microarray biclustering algorithms include Kluger *et al.*'s spectral model [66] and Lazzeroni *et al.*'s plaid model [68].

While many biclustering algorithms have been developed for gene microarray analyses, one of the first applications for biclustering was text mining. Dhillon *et al.* proposed two different biclustering algorithms for simultaneously partitioning documents and words: spectral co-clustering [24], and a co-clustering algorithm based on information theory [25]. Kluger *et al.*'s spectral model [66] for gene microarray analyses is based on Dhillon *et al.*'s spectral model [24].

More recently, biclustering has been applied to single cell RNA sequencing (scRNA seq) data. Biclustering methods specific to this application include BackSPIN [136] and QUBIC2 [131].

Rugeles *et al.* [98] developed Dual Topics for Bicluster (DT2B), a biclustering method based on a generalized latent Dirichlet allocation (LDA) model [8]. Unlike the previous models, DT2B avoids the constraints of a model structure. However, the algorithm requires a discretized data set, pre-specification of the number of row and column clusters, and threshold values.

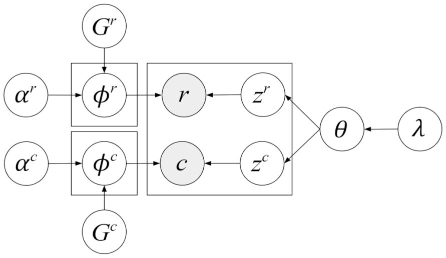
By using a DPMM instead of a LDA model, we bypass the need to specify the number of biclusters, make strong modeling assumptions, and particular data format.

3.4 Background

3.4.1 Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a hierarchical Bayesian model used to infer latent features in collections of discrete data. Initially proposed to estimate and describe population structure from genotype data [91], it is also commonly used for the classification of documents based on word frequencies [8].

In the context of document classification, LDA posits that for a corpus of documents, the probability distribution of words for a given document is determined by a set of latent "topics" associated with that document. LDA infers these latent topics from observed word frequencies for each document to produce a clustering or classification of documents in the corpus.



Variable	Distribution	Description
r_i	$Cat_\infty(\phi_r)$	Observed row
c_i	$Cat_\infty(\phi_c)$	Observed column
(z_i^r, z_i^c)	$Cat_{\infty\times\infty}(\theta)$	Row and col. latent variables
θ	$Dirichlet(\lambda \vert z^r, z^c)$	Joint latent variable distribution
ϕ^r	$DP(\alpha^r)$	Row latent variable distribution
ϕ^c	$DP(\alpha^c)$	Col latent variable distribution
α^r, α^c	–	DP hyperparameters
λ	–	Hyperparameter for θ

Figure 3.1: Plate diagram and description of included variables and parameters. Rows r and columns c are clustered separately through the DPMMs defined by ϕ^r and ϕ^c. After updating according to Algorithm 1, heavy biclusters can be extracted from ϕ^r, ϕ^c, and θ (the joint distribution of latent cluster assignments given by z^r, z^c).

3.4.2 Parallel Sampling of DPMMs

To review the explanation of Dirichlet process mixture models (DPMMs), please refer to Chapter 2.4.3. DPMMs have been largely computationally heavy to implement. [15] parallelized the MCMC sampler for DPMMs by utilizing a restricted Gibbs sampler to fix the number of clusters before proposing splits and merges. Since the number of clusters are fixed, each of the Gibbs sampler steps can be done in parallel. Furthermore, to increase efficient cluster splits, they augment each cluster with two sub-clusters, labeled $\bar{z}_i \in \{l, r\}$ to denote whether each data point x_i is associated with the left or right sub-cluster. Additional auxiliary variables introduced are the sub-cluster weights $\bar{\pi}_k \in \{\bar{\pi}_{k,l}, \bar{\pi}_{k,r}\}$ and parameters $\bar{\theta}_k \in \{\bar{\theta}_{k,l}, \bar{\theta}_{k,r}\}$ of cluster k. The auxiliary variables for the sub-clusters are analogous in function to the variables for the regular clusters. In this augmented restricted Gibbs sampling algorithm, we now sample a regular cluster assignment and then a sub-cluster assignment for each data point. Splits and merges, to either split a cluster into its two sub-clusters or merge two sub-clusters into one new cluster, are proposed and accepted with probability $\min(1, H)$, where $H \in \{H_{split}, H_{merge}\}$ is the Hastings ratio for the respective action.

[27] extended this implementation to enable parallelization on multiple multi-core machines instead of a single multi-core machine. The authors note that sampling cluster parameters θ_k is parallelizable over the clusters, sampling cluster assignments z_i is independently com-

puted for each data point x_i, and proposing cluster splits is parallelizable. For computational efficiency, they rely on a distributed-memory model and utilize sufficient statistics to communicate between the cores as well as the between the machines. The sufficient statistic T for a multinomial cluster (e.g. for document classification or single cell RNA sequencing data analysis) is $T = \sum_{i=1}^{N} x_i \in N_0^d$, where d is the dimension of the data points x_i. The aggregation of the sufficient statistics for each cluster allows for the sampling of cluster parameters across multiple parallelized worker processes. Splits and merges are proposed similarly to [15] on the master process, with mappings of old cluster assignments to new assignments broadcasted to all worker processes to individually update its data points. Using this multi-machine, multi-core implementation considerably speeds up our model and allows us to handle high dimensional data.

3.5 Conjoined Dirichlet Process (CDP)

CDP is a probabilistic biclustering method that provides several important characteristics in the context of gene-cell count analysis. The estimated biclusters may overlap, posterior probabilities of each element belonging to a given bicluster can be calculated, and heavy biclusters (showing strong co-occurrence in rows and columns) are encouraged. By utilizing a pair of DPMMs for bicluster estimation, CDP eliminates the need to specify the number of row topics and columns topics *á priori*, which is particularly relevant for both gene expression and document analysis where the number of topics and biclusters is unknown or ill-defined.

3.5.1 Model Construction

CDP can be summarized in two steps:

1. Use DPMMs to learn row and column clusters.

2. Model the mutual dependence between the row and column clusters to extract biclusters with strong co-occurrence values in both rows and columns.

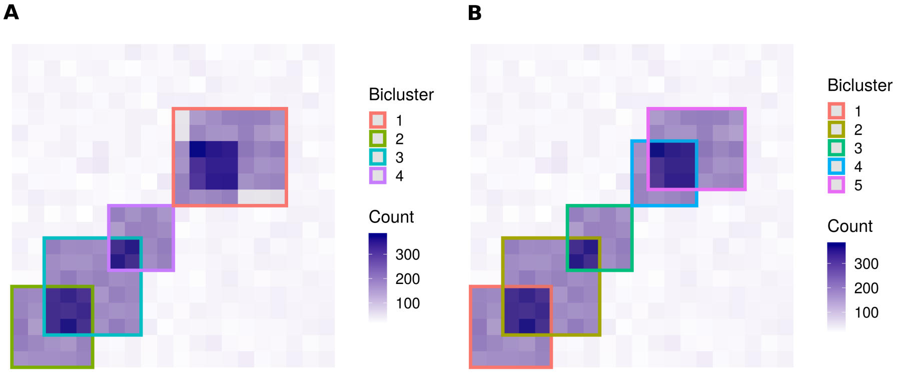


Figure 3.2: CDP is able to detect overlapping biclusters: (A) Heatmap of simulated count data and bicluster membership estimated by CDP. (B) True bicluster structure for simulated data.

Given a $n_R \times n_C$ matrix where n_R is the number of rows and n_C is the number of columns, each matrix entry (r, c) represents the frequency of row r in column c. For text data, this corresponds to the frequency of word r in document c and for single cell RNA sequencing data, this corresponds to the gene expression of gene r in cell c.

Using a DPMM, we can sequentially cluster the rows and columns of the matrix to obtain row-cluster assignments z^r and column-cluster assignments z^c. Similar to DT2B [98], we now have two sets of latent variables (e.g. topics for text data) and use these sets to extract biclusters with strong co-occurrence values in rows and columns.

Figure 3.1 shows the graphical model for CDP, where row r and column c are the rows and columns of the data matrix. z^r and z^c are the vectors of row and column cluster indices (assignments) respectively. ϕ^r is the row per row latent variable distribution, ϕ^c is the column per column latent variable distribution, and θ is the joint latent variable distribution. These three variables maintain the counting over the relationships between the data, latent variables and their mutual dependence. For discrete data, the hyperparameters for CDP are γ, the concentration parameter for the DP; β, the prior for the DP measure; α^r, the hyperparameter for ϕ^r; α^c, the hyperparameter for ϕ^c; and λ, the hyperparameter for θ. Figure 3.2 shows an illustrative example of CDP.

Theorem 3.1. *If row assignments z^r are held fixed, then the CDP update step is equivalent to a latent Dirichlet allocation update on z^c. A similar result holds if z^c is held fixed for updating z^r.*

Proof: In evaluating Eq. 3.4 to update z^c, we can then treat ϕ^r as a constant, yielding

$$
\begin{aligned}
P(z_i^c = j | c_i = m, r_i = n, \mathbf{z}_{-i}^c) && (3.1)\\
\propto \phi_{mj}^c \theta_{jk} && (3.2)\\
\propto \frac{C_{mj} + \alpha^c}{\sum_{m'} C_{m'i} + n_C \alpha^c}(C_{ij} + \lambda). && (3.3)
\end{aligned}
$$

Updating z^c according to this probability is equivalent to the update given by LDA [8].

3.5.2 Inference Process

Algorithm 1 shows the inference process for CDP, using the distributed MCMC inference algorithm outlined in [27], which is based on the restricted Gibbs sampler method in [15]. Due to the split and merge aspect and the high-dimensionality of our data, the posterior distribution of the assignment parameters may be multi-modal. For this reason, we update the assignment parameters for a specified number of iterations and take the MAP estimate of the maximum values of z^r and z^c.

The specifications for the hyperparameters of CDP (under the assumption that the base distribution of the DP is multinomial) are listed below:

$$\begin{aligned} \gamma^r, \gamma^c &\in R^{1\times 1} & \alpha^r &\in R^{K_r\times 1}\\ \beta^r &\in R^{n_R\times 1} & \alpha^c &\in R^{K_c\times 1}\\ \beta^c &\in R^{n_C\times 1} & \lambda &\in R^{K_r\times K_c} \end{aligned}$$

Given a collection of composites (e.g. documents, cells) C made up of parts (e.g. words, genes) R, we can write the probability of a composite c containing a part r as:

$$P(r, c) = \sum_{z_r} \sum_{z_c} P(r|\phi^r_{z_r}, \alpha^r) P(c|\phi^c_{z_c}, \alpha^c) P(z_r, z_c|\theta)$$

A major advantage of the CDP over DT2B [98] is that the CDP does not require thresholds to control the trade-off between quantity and quality of the biclusters. The hyperparameters in the DPMM step facilitate this trade-off automatically. Setting a large γ, the Dirichlet process concentration parameter, and for a multinomial base distribution, a large β (the Dirichlet distribution hyperparameter) will yield more clusters.

The probabilistic biclusters are given by the joint distribution of row and column latent variables, θ, which has dimension $K_r \times K_c$. K_r and K_c are the number of latent row and column variables respectively. As previously mentioned, from the DPMMs, we obtain the row-cluster assignments z^r and column-cluster assignments z^c. Calculating the mode of the posterior distributions of z^r and z^c yields the maximum a posteriori (MAP) estimate of the number of latent row and column variables, i.e. K_r and K_c.

We note that the dimensions of the row per row latent variable distribution, ϕ_r, and the column per column latent variable distribution, ϕ_c, are also given by the MAP. ϕ_r has dimensions $n_R \times K_r$, and ϕ_c has dimensions $n_C \times K_c$.

Algorithm 1 Conjoined Dirichlet Process (CDP)

```
Input: Data $X$, size $n_R \times n_C$
       DP concentration parameters $\gamma_R, \gamma_C$
       Dirichlet distribution hyperpriors $\beta_R, \beta_C$
       Number of DPMM iterations $iter_R$, $iter_C$
       Number of cluster reassignment iterations $iter_U$
for $i = 1$ to $iter_R$ do
  Run DPMM on $X$
end for
for $i = 1$ to $iter_C$ do
  Run DPMM on $X^T$
end for
Calculate $K_r = MAP(z^r)$ and $K_c = MAP(z^c)$
for $i = 1$ to $iter_U$ do
  Update $z^r$ and $z^c$ using the data as weights
end for
for $i = 1$ to $n_R$ do
  for $j = 1$ to $K_r$ do
    Calculate $\phi^r_{ij} = \frac{C_{ij}+\alpha^r}{\sum_{i'} C_{i'j}+n_R\alpha^r}$
  end for
end for
for $i = 1$ to $n_C$ do
  for $j = 1$ to $K_c$ do
    Calculate $\phi^c_{ij} = \frac{C_{ij}+\alpha^c}{\sum_{i'} C_{i'j}+n_C\alpha^c}$
  end for
end for
Calculate $\theta \propto C_{r,c} + \lambda$, $1 \leq r \leq n_R, 1 \leq c \leq n_C$
```

3.5.3 Bicluster Extraction

From the DPMM, we obtain latent variables z^r and z^c, which indicate the row and column cluster assignments respectively. To extract the biclusters from the data, we need to calculate three parameters: row per row latent variable distribution ϕ^r, column per column latent variable distribution ϕ^c, and joint distribution of row and column latent variables θ.

These three parameters are given by [98]:

$$\phi^c_{mi} = \frac{C_{mi} + \alpha^c}{\sum_{m'} C_{m'i} + n_C \alpha^c} \tag{3.4}$$

$$\phi^r_{nj} = \frac{C_{nj} + \alpha^r}{\sum_{n'} C_{n'j} + n_R \alpha^r} \tag{3.5}$$

$$\theta \propto C_{ij} + \lambda \tag{3.6}$$

where C_{ab} is the number of instances a-th variable is assigned to b-th variable. For example, ϕ^r is the probability of the n-th row being assigned to j-th row latent variable. Thus, C_{nj} is the number of times the n-th row is assigned to to j-th row latent variable. The joint distribution θ tracks the relationship between the current row and column latent variables to capture the mutual dependence between the two sets of latent variables [98].

First, we calculate ϕ^r and ϕ^c by using the aforementioned sets of latent variables z^r and z^c as the initial cluster assignments. These assignments are updated iteratively using the data as weights. Once the row and column assignments have been updated, we count the number of instances a row or column is assigned to a row or column latent variable.

To obtain the joint distribution of row and column latent variables θ, we need to calculate the frequency of each row and column latent variable pairing (i, j). The vector of frequencies for each row and column latent variable pairing is then transformed into a contingency table of size $K_r \times K_c$, i.e. the desired θ.

3.5.4 Implementation Overview

To obtain the row-cluster assignment z^r and column-cluster assignment z^c, we separately infer each parameter using [27]'s implementation in Julia. We utilize a specific version of that package that outputs the cluster assignments z^r or z^c at each iteration rather than the final cluster assignments. While this requires more memory storage and run time, it allows CDP to have overlapping biclusters and more interpretable results depending on the application.

For N data observations, K clusters, and M machines with P cores, the total runtime complexity for the DPMM implementation is $\mathcal{O}(K)+\mathcal{O}(M+P)+\mathcal{O}(NK/(MP))$. For more details on the runtime complexity for the DPMM, see [27].

CDP reassigns each observation iteratively in batches of size equal to either the row sums or column sums. These batches are parallelized to run on P processes (cores). Thus, updating the row and cluster assignments for J iterations takes $\mathcal{O}(NJ/P)$ time. Calculating ϕ for the rows and columns require the aforementioned assignment step. Once reassigned, CDP splits the N data points into vectors of length row sums (for ϕ^r) or column sums (for ϕ^c). These vectors are then tabulated over the number of latent variables K to determine the probability of each row or column being assigned to each row latent variable or column latent variable respectively. The runtime complexity for calculating ϕ excluding the cluster assignment update step is then $\mathcal{O}(NK)$ where K is equal to K_r when calculating ϕ^r and K_c when calculating ϕ^c. Calculating the joint distribution of both row and column latent variables θ requires looping over the assignments for one direction (e.g. row assignments) and matching the row and column indexes to the assignments in the other direction (e.g. column assignments). This operation requires $\mathcal{O}(N)$ time. CDP then tabulates the row and column assignment of each row and column pairing to obtain θ. Thus, the total runtime complexity for calculating θ is $\mathcal{O}(NK_rK_c)$ and $\mathcal{O}(NK_rK_c/P)$ if run in parallel.

As $N \gg K, P, M$ and J, CDP takes $\mathcal{O}(NJ)+\mathcal{O}(NK_rK_c)$ time. Experiments were conducted on an i5-7600K CPU.

3.6 Experimental Results

We compare CDP to DT2B [98] because this method also models the mutual dependency between two sets of latent variables. We also compare our algorithm to spectral biclustering [66] since both try to extract high co-occurences. For completeness, Cheng and Church [17] and the plaid [68] algorithms are also used for comparisons due to their common usage, and BiMax [90] which is known to serve as a reference method.

3.6.1 Synthetic Data

Simulated count data were generated from a multinomial distribution defined by an $R \times C$ probability matrix θ (with entries summing to 1), and by fixing the sum of entries in the resulting random matrix at some total count N. The total bicluster probability p of an element belonging to a bicluster was set to control the strength of biclusters and overall sparsity. Four different constructions of θ were chosen to evaluate performance over different biclustering patterns. In order of increasing complexity, these four cases are (1) a single distinct bicluster, $N = 4000, R = C = 50, p = 0.8$; (2) two distinct biclusters, $N = 4000, R = C = 20, p = 0.5$; (3) 3 biclusters with one overlap $N = 4000, R = C = 50, p = 0.7$; (4) 5 distinct biclusters, $N = 10000, R = C = 100, p = 0.7$ (see Figure 3.3 for an example).

To compare the performance of CDP to existing methods, we use the Jaccard score, defined as $J(\mathcal{B}_1, \mathcal{B}_2) = \min_{(\mathcal{A},\mathcal{B})} \frac{1}{|\mathcal{A}|} \sum_{A \in \mathcal{A}} \max_{B \in \mathcal{B}} \frac{|A \cap B|}{|A \cup B|}$, where $\mathcal{B}_1, \mathcal{B}_2$ are two sets of biclusters, with the minimum taken over $(\mathcal{A}, \mathcal{B}) \in \{(\mathcal{B}_1, \mathcal{B}_2), (\mathcal{B}_2, \mathcal{B}_1)\}$. The Jaccard score is a symmetric similarity metric taking values $0 \leq J(\mathcal{B}_1, \mathcal{B}_2) \leq 1$, with the lower bound attained only when all sets in $\mathcal{B}_1$ are disjoint with all sets in $\mathcal{B}_2$ and the upper bound attained only when $\mathcal{B}_1 = \mathcal{B}_2$. In the context of the simulation study, $\mathcal{B}_1$ is the set of estimated biclusters from a given method, and $\mathcal{B}_2$ is the set of true biclusters from the generative model.

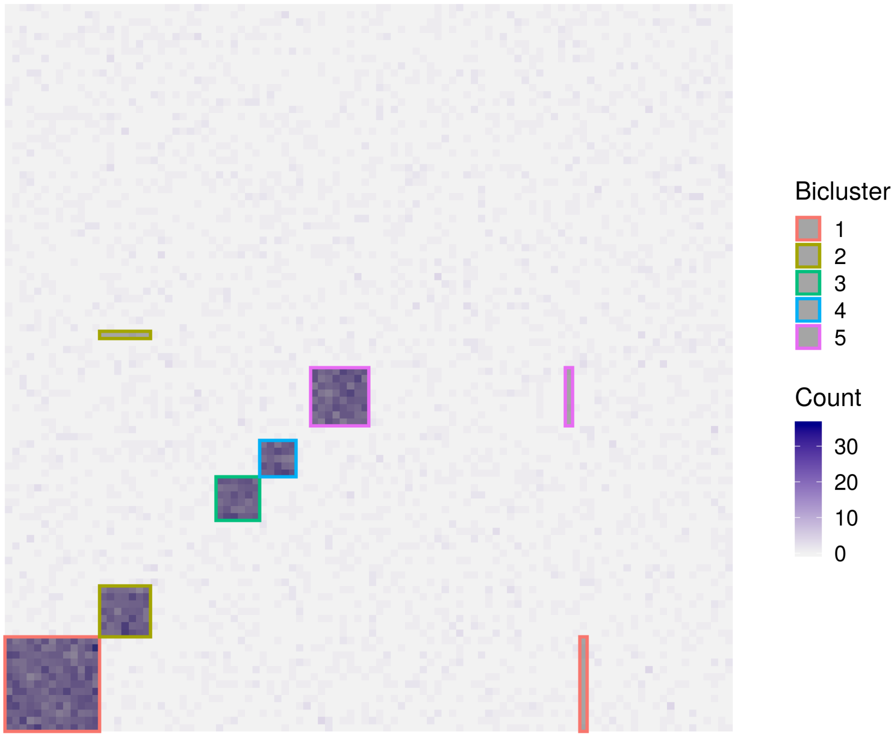


Figure 3.3: Example results from CDP for simulated data (case 4). CDP correctly identifies the heavy biclusters (0.938 Jaccard score), with only a small number of spurious elements included (e.g. in biclusters 1, 2 and 5). The data shown here is approximately 70% sparse.

3.6.2 Real-life Data

1. **Condensed 20 Newsgroups**: Collection of 100 words across 16,242 newsgroup documents (”netnews”). The data is organized into 17 different newsgroups and 4 main topics. This data set is 95.97% sparse.

2. **Single cell RNA sequencing (scRNA seq) Data**: Collection of 23,226 genes across 5,053 transcriptomes from 10 distinct regions of murine juvenile and adult central nervous system [74]. All cells were profiled using the Fluidigm C1 system and sequenced on an Illumina HiSeq 2000 instrument. This data set is 87.57% sparse.

3.6.3 Parameter Settings

For CDP, we need to set the number of iterations, the Dirichlet process concentration parameter γ, and the Dirichlet distribution hyperprior β. Note that both our text and biological data are discrete counts so we assume a multinomial base distribution. If we had continuous data we would instead assume a Gaussian base distribution (or another continuous distribution) and set the values for a Normal–Inverse–Wishart hyperprior. The hyperparameters for ϕ^r, ϕ^c and θ are set to zero by default. We set all concentration parameters and hyperpriors to be small to obtain larger cluster sizes. Table 3.1 shows the parameter values for the two real data sets. We did not include the two β hyperparameters or the λ hyperparameter in the table since we set those values to zero. In practice, if one has strong prior knowledge regarding a row or column element, setting a value greater than zero for those hyperparameters will result in a more accurate clustering. However, we are doing strictly exploratory work for this paper.

Table 3.1: Parameter settings for the DPMM part of CDP on two data sets.

Data set	Row/Col	Iterations	γ	β
Newsgroups	Row	1000	10	1
	Col	1000	100	1
scRNA Seq	Row	500	10	0.1
	Col	500	10	1

3.6.4 Results for Synthetic Data

Results for the four synthetic data cases are provided in Table 3.2. In each of the cases considered, plaid, DT2B, and CDP exhibit the highest accuracy in bicluster estimation as

Table 3.2: Comparison of mean (standard deviation) Jaccard similarity scores for various biclustering algorithms on the simulated data sets.

Case	Plaid	C & C	Bimax	DT2B	CDP
1	0.133 (0.06)	0.16 (0.00)	0.141 (0.005)	0.806 (0.204)	0.69 (0.088)
2	0.862 (0.265)	0.016 (0.000)	0.098 (0.021)	0.889 (0.118)	0.985 (0.081)
3	0.172 (0.087)	0.048 (0.000)	0.105 (0.012)	0.236 (0.032)	0.25 (0.014)
4	0.316 (0.343)	0.008 (0.000)	0.101 (0.014)	0.522 (0.087)	0.756 (0.033)

measured by the Jaccard score. We also tested the spectral method, but the accuracy was so low we excluded it from the table. In cases 2, 3, and 4, CDP outperforms all other methods, and gives substantially better performance in the most complicated setting (case 4), with a mean Jaccard similarity of 0.756, compared to DT2B with a mean score of 0.522. CDP also exhibits lower variance over repeated simulations compared to DT2B. Only in the simplest setting of a single bicluster (case 1) does DT2B show better mean similarity score, with 0.806 for DT2B compared to 0.69 for CDP. However, DT2B shows high variance in the accuracy of its estimates over repeated runs in this case, whereas CDP shows lower variance over all scenarios. Together, these results suggest that CDP is practical for bicluster extraction, and may be significantly more accurate compared to existing methods.

3.6.5 Simulation Runtime Comparisons

The DT2B method is conceptually similar to CDP, but requires selecting a maximum number of row and column clusters to determine the parameters of the underlying LDA models. In general, DT2B is most efficient and accurate when the maximum number of row clusters (K_r) and column clusters (K_c) are set to the true number of row and column clusters respectively, but these values will be unknown in practice. DT2B runs in $O(NK_rK_c)$ time [98], thus setting K_r and K_c to the number of rows and columns respectively may be computationally prohibitive for applications to single cell analysis and other large data settings. Table 3.3 shows the runtime of DT2B for different choices of K_r and K_c on a simulated data set, compared to CDP.

Table 3.3: Comparison of runtimes for CDP and DT2B with various choices of (K_r, K_c) on a simulated data set (case 2). Runtimes for DT2B scale linearly in both K_r and K_c.

Method	Mean Jaccard (s.d.)	Runtime (s)
CDP	0.96 (0.02)	13.22
DT2B(5, 5)	0.41 (0.11)	4.23
DT2B(10, 10)	0.94 (0.13)	12.85
DT2B(25, 25)	0.98 (0.01)	73.45

3.6.6 Results for Text Data

Biclustering text data from a document corpus allows for identification of document-word combinations with high co-occurrence. Extracted biclusters represent combinations of words and documents that form a (latent) topic. This is distinguished from traditional LDA topic modeling in that LDA does not cluster documents directly, and words which co-occur across many documents may be clustered even if the shared vocabulary of those documents is small overall. Instead, a biclustering such as CDP encourages heavy topics which exhibit high co-occurrence of words across documents and documents across words.

The condensed version of the 20 Newsgroup data set is organized into 17 different newsgroups corresponding to four main topics: comp (e.g. computing, graphics), rec (e.g. recreational, sports), sci (e.g. medicine, electronics, space) and talk (e.g. politics, guns), and two smaller topics: religion and miscellaneous for sale.

CDP found 5 word clusters, 3 news groups, and 3 heavy biclusters. There is generally no ground truth for biclusters on text data, and due to the overlapping nature of this ”netnews” data set, we chose to evaluate the biclusters by visual inspection. We present Table 3.4 showing the words with the highest co-occurrences across documents. The first grouping is predominantly about space and political topics, while the second grouping is comprised of recreational, religious and medical topics. The third heavy bicluster consists of computational topics.

Table 3.4: Selection of the top six words with the highest co-occurrence values across the documents.

Topic 1	Topic 2	Topic 3
Mars	children	ftp
solar	disease	fans
technology	bible	files
satellite	baseball	format
shuttle	cancer	fact
president	patients	games

3.6.7 Results for Single cell RNA Sequencing Data

Biclustering scRNA seq data is commonly used to define developmental stages based solely on the transcriptome in addition to accounting for variation in the data, and identifying biologically important genes and their signatures for each cell stage. Each bicluster is an association between groups of cell stages and their genetic drivers.

A key contribution of CDP is the ability to identify the cell stages and their genetic drivers without having to find highly expressed genes *á priori*. Furthermore, cell stages are dynamic in time and a probabilistic clustering assignment allows us to capture part of this dynamic without a true time series model. This contribution is a vital reason as to why we utilize the MAP to determine the most probable number of clusters instead of running the two DPMMS until they converge on a single value.

We apply CDP to the scRNA seq data set in [74]. The authors performed a biclustering analysis using BackSPIN [136] and found 13 cell clusters.

CDP found 7 gene clusters, 12 cell clusters, and 4 strong biclusters. Like text data, there is generally no ground truth for biclusters on scRNA seq data. We evaluate our method using the PANTHER classification system and tools [78, 114, 79], and also compare it to [74]'s results.

The four biclusters with the strongest co-occurrence values consist of myelin-forming oligodendrocytes (MFOL2), and several stages of mature oligodendrocytes (MOL5, MOL4 and MOL3). Biclusters with weaker co-occurrence values consist of newly formed oligodendrocytes (NFOL1) and oligodendrocyte precursor cells (OPC). The oligodendrocyte precursor cells can differentiate into newly formed oligodendrocytes, which produce myelin and continue maturing. Since there are multiple stages of maturation, the composition of the strong biclusters are expected and are corroborated by [74]. The majority of the oligodendrocyte cells are no longer precursor cells or newly formed; they are in differing stages of maturation.

Furthermore, CDP shows that the oligodendrocyte classes also correspond to different regions of the central nervous system. For example, oligodendrocytes classified as MFOL2 are also found in abundance in the substantia nigra ventral tegmental (SN-VTA) and hypothalamus regions of the central nervous system. Likewise, oligodendrocytes classified as MOL5 are found in abundance in the dorsal horn.

With respect to the genes, CDP did not find distinct gene groupings. However, CDP did find two overlapping groupings and multiple groupings with weak co-occurrence values. Using PANTHER, we find that the two overlapping groupings are strongly affiliated with binding, particularly enzymatic binding, and catalytic activity. One group is more involved with cytoskeletal protein binding, and at a higher cellular level, is associated with cellular response to stimulus and cellular metabolic processes. The second group is more involved with signaling receptor binding, and with cell component organization and signal transduction at a higher level. Genes associated with other biological processes such as the lipid metabolic process or the multicellular organismal process are in the biclusters with weaker co-occurrence values.

3.7 Discussion

In this paper, we presented a novel, non-parametric probabilistic biclustering method designed to address the challenges of model and parameter selection required by competing methods. By utilizing two infinite mixture models and calculating their mutual dependence, we are able to estimate the number of biclusters strictly from the data and prior, and identify the biclusters without strong modeling assumptions.

CDP currently requires hyperparameter specifications, but putting a prior on these hyperparameters may improve accuracy without the need for running the model over a range of parameters. Furthermore, CDP is focused on partitioning discrete data since text and scRNA seq data naturally have count data. However, other applications such as audio retrieval do not. CDP has the ability to model continuous data as well by changing the multinomial base distribution to a Normal–Inverse–Wishart base distribution and modifying the mutual dependence calculation steps.

Simulation results suggest CDP significantly improves upon DT2B and current standard methods, with more accurate estimation of biclusters, and lower variance estimates. Experimental results on real data with high sparsity ($> 85\%$) demonstrate that CDP is able to extract meaningful heavy biclusters. In single cell analyses, this advantage is particularly useful as the data is extremely sparse and noisy. However, CDP is computationally expensive compared to other clustering methods and is limited to smaller single cell data sets. This is not ideal as the size of single cell data sets are increasing as sequencing technology develops.

As a probabilistic model leveraging DPMMs for bicluster estimation, CDP can easily be extended to include additional structure and assumptions. For instance, in the context of single cell analysis, known results on gene networks may be incorporated through the DPMM priors. Furthermore, by choosing continuous DPMM base measures G^r, G^c, CDP can be applied for biclustering a matrix of continuous values, providing an important advantage over DT2B, which can only accommodate discrete values.

3.8 Data and Software

All data sets are publicly available. The condensed 20 Newsgroup data set is available on Sam Roweis's website [97]. The scRNA seq data set is part of the Hemberg lab's collection of publicly available scRNA seq data sets [64] as a SingleCellExperiment Bioconductor S4 class [72]. We removed rows and columns where the entire vector consisted of zeros. For the scRNA seq data set, we also combined the counts of genes that had been split into multiple entries based on loci position.

Source code for CDP can be found at `https://github.com/micnngo/CDP`. DPMMs were run using the 'exposed_parr' branch of DPMMSubClusters [27]. The main CDP script is written in R with a wrapper for Julia and C++. Plaid, Cheng and Church, Spectral and BiMax algorithms were run using the package 'biclust' in R [58]. The source code for DT2B is available on Github [98].

3.9 Acknowledgements

This work was supported by NSF grants DMS1936833 and DMS1763272, Simons Foundation grant (594598, QN), and NIH grant R01MH115697. We also thank Or Dinari for sharing a modification to his DPMMSubClusters package.

Chapter 4

Leveraging information from groups of genes to identify cell groups

4.1 Introduction

In Chapter 2, we introduced an experiment to investigate how TNF-α affects a normal patient versus a MPN patient. We identified differentially expressed genes between the unstimulated samples (not treated with TNF-α) and the samples treated with TNF-α in each of the two patients. Then in Chapter 3, we introduced a method titled "Conjoined Dirichlet Process" to simultaneously cluster the genes and cells. While CDP worked well on synthetic data and smaller data sets, we were not able to use CDP to analyze the experimental data of interest and obtain meaningful results. We are still interested in investigating the differentially expressed genes between the unstimulated and treated samples among the different cell populations in each of the two patients.

Recently, in contrast to relying solely on the highly variable genes to cluster the cells, several methods have proposed alternative approaches. For example, Vandenbon and Diez's singleCellHaystack [118] uses the Kullback-Leibler divergence to find cell type marker genes independent of any prior clustering; Qiu leverages the dropouts to perform co-clustering to identify cell types [93]; Zhang *et al.*'s Single-cEll Variable gEnes (SIEVE) [139] utilizes random sampling for all single cells in a data set to produce a more robust set of highly variable genes; and Kim *et al.*'s MarcoPolo [63] identifies informative marker genes independent of any prior clustering using a three-step approach based on a bimodal mixture model, voting system and whether the cells expressing a candidate gene are proximal to each other in a low-dimensional space. To this end, since we have already obtained ranked groups of differentially expressed genes in our data sets of interest in Chapter 2, we approach the identification of homogeneous cell subpopulations in primary human bone marrow and hematopoietic stem and progenitor cells without explicitly selecting for highly variable genes.

As we have noted in the previous chapters, while scRNA seq data can use tools developed for bulk RNA seq data and microarray data, its noise and dimensionality largely prohibits using any of the more classic biclustering algorithms to co-cluster the data. Furthermore, feature selection is a key pre-processing step in scRNA seq data. The noise from including all the genes and cells that passed quality control would muddle the biological significance and obfuscate the results. Thus, we propose a sequential biclustering pipeline that leverages the expression from the ranked groups of genes identified in Chapter 2.

4.2 Method

4.2.1 Pre-processing

For input into our sequential biclustering pipeline, we use the output from the differentially expressed pipeline described in Chapter 2.4 and build on the pre-processed data sets from Chapter 2.4.1. As a recap, we filtered out low quality cells and any genes where there are no expression in both the untreated and stimulated samples. We then normalized the count data matrices and log-transformed them to yield two log-normalized data sets for each patient. The output from Chapter 2 is a data frame containing the gene names, the W statistics, p-values, scaled z-scores and cluster IDs, which are ranked from insignificant to significant (1 to K the number of clusters). From the output, we select all the gene groups containing scaled z-score values with a magnitude greater than 2 and subset each of the log-normalized data sets to include only these "significant" gene groups. The rest are discarded from further analysis. These significant gene groups will be used downstream to create a gene latent space to cluster the cells.

For the rest of the pipeline, we require one data matrix so we simply merge the unstimulated data set with the stimulated data set. Note that we have ensured that there is no batch effect from the stimulation status, so the resulting clustering from this merged data set should not be trivial where one cluster is the unstimulated group and the other is the stimulated group. If there is a batch effect from the merging, we suggest integrating the unstimulated and stimulated data sets using methods such as Seurat's [46, 108] `IntegrateData` function instead of merging the two data sets.

4.2.2 Creating the gene latent space

Since each gene grouping should contain only genes with similar latent expression profiles, we can reduce the dimension of the pre-processed gene expression matrix by reducing the dimension of each gene grouping to 1. For example, if the results from Chapter 2 yielded n_{gg} significant gene groups and there are n_{cells}, the reduced gene expression matrix X^G will have dimensions $n_{cells} \times n_{gg}$. Although more latent dimensions can be kept for each gene group, we will concatenate all latent dimensions to create a gene latent space matrix. Keeping more than one latent dimension per gene group introduces potential dependency which may affect downstream analyses.

To reduce the dimensionality of each gene grouping, we can use most standard approaches to capture the information. As a full proof of concept, we try PCA, t-SNE and UMAP to reduce the dimensionality of each gene grouping and keep only the first latent dimension for each group. Since we will be using each gene group's latent representation as covariates for the cells, keeping only the first latent dimension to reduce high correlation between the covariates. Figures 4.1 and 4.2 shows the latent spaces for each gene grouping if we had kept two latent spaces per group.

4.2.3 Creating the cell latent space

Once we reduce each gene group to one latent dimension, we can concatenate these latent dimensions to obtain our cell latent space. Given the gene latent space matrix X^G with dimensions $n_{cells} \times n_{gg}$, we calculate the Pearson correlation matrix to determine which cells have a stronger relationship with each other given its gene latent space expression profiles. A value of zero indicates no linear relationship between the cells and values of 1 or -1 indicate a strong positive linear relationship (correlation) or negative linear relationship respectively. However, with dimensions $n_{cells} \times n_{cells}$, its size creates computational challenges.

To mitigate this issue, we try to remove "redundant" information by reducing the correlation matrix to two dimensions. This yields what we call the cell latent space expression matrix X^C with dimensions $n_{cells} \times 2$. Due to the size of the correlation matrix, the two methods of dimensionality reduction that work the best in terms of efficiency and biological significance appear to be UMAP and an autoencoder. For this chapter, we reduced the dimensions of the correlation matrix for the unaffected patient using UMAP and for the MPN patient using an autoencoder with 3 hidden layers (256-, 64-, and 2-units).

4.2.4 Cluster the cells

In the previous sections, we created a reduced expression matrix X^C with dimensions $n_{cells} \times 2$ that encapsulates the gene expressions of each cell. With this relatively "small" matrix size, arguably, most conventional clustering algorithms will be sufficient. To cluster X^C, we use a fairly simple approach: k-means clustering. Using this and the within-sum-of-squares (WSS) to figure out the "best" number of clusters, we can obtain k clusters. For a non-parametric approach, we can use a Dirichlet process mixture model (briefly described in Section 2.4.3) with $\alpha = 0.1$ and obtain a data-driven optimal number of clusters. We note that we can modify the cluster resolution accordingly and further subdivide or combine clusters after performing further downstream analyses to determine each cluster's biological relevance.

4.2.5 Post-processing

To examine biological relevance, we must perform differential gene expression analysis to label each cell cluster as clustering only groups the cells. Here, we again perform the Wilcoxon test described in Section 2.4.2 but instead of comparing the stimulated sample against the unstimulated sample, we will be comparing one cell cluster against all the other cell clusters. Since we have k cell clusters, we will repeat this process k times to obtain a set of differential genes ("marker") genes for each cluster.

4.3 Results

4.3.1 Clustering results: Unaffected Patient

From the pipeline outlined in Chapter 4, we obtain roughly seven cell clusters. Figure 4.1 shows the gene and cell latent representations. The top row of the figure visualizes the first two latent dimensions of each gene cluster. We note that there are no batch effects here with respect to their stimulation status, and that the differences between the three dimensionality reduction methods are more striking between the UMAP aprroach versus the other two. All three approaches suggest there may be at least two main gene clusters.

In Section 4.3.3, we will also examine the marker genes for each of these identified cluster to determine what each group is.

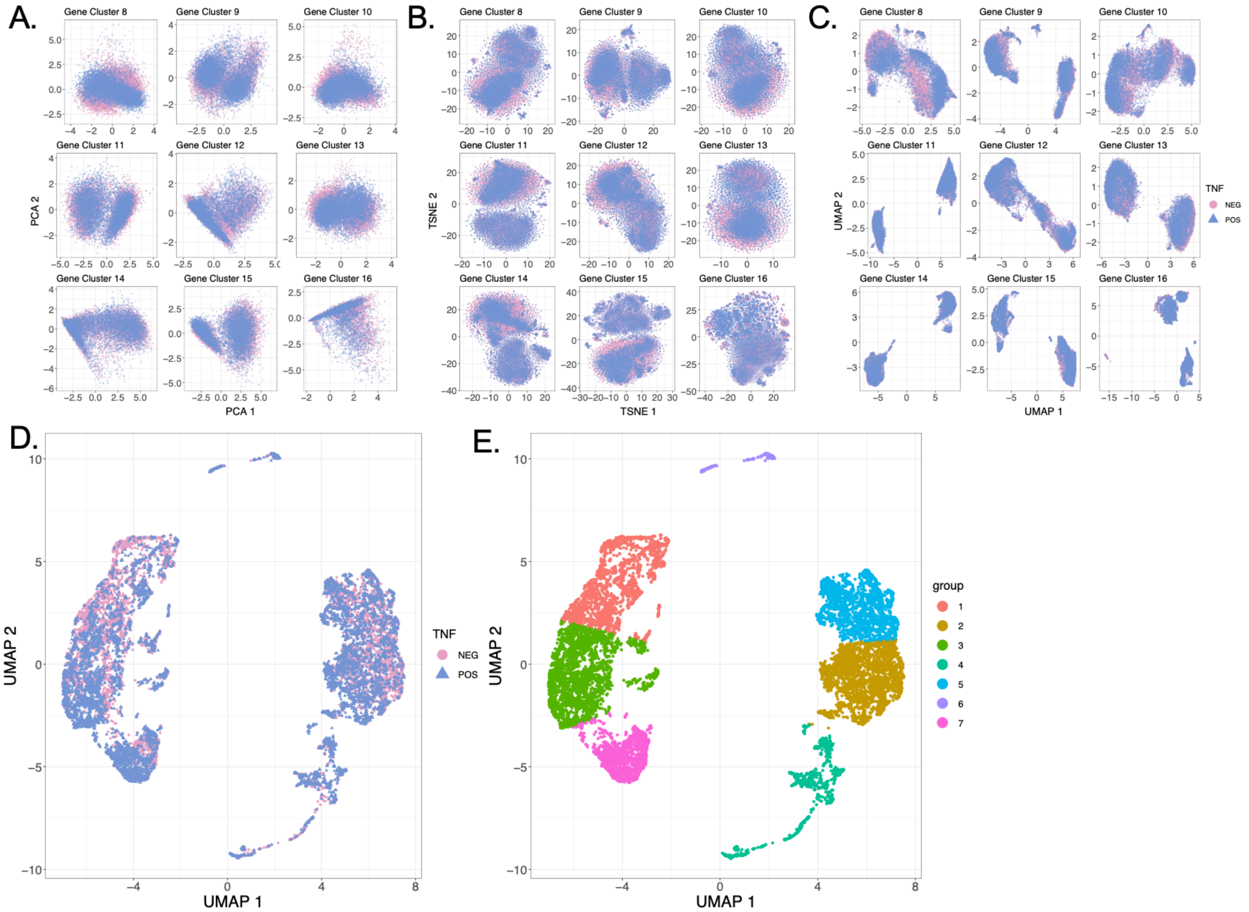


Figure 4.1: We visualize the latent representation of each gene cluster using PCA (A), *t*-SNE (B) and UMAP (C) for the unaffected patient. After we obtain the gene latent representation, we can calculate the cell latent representation (D) and then cluster the cells (E).

4.3.2 Clustering results: MPN Patient

In the MPN patient, we obtain roughly 3 main cell clusters but further subdivide to yield 7 cell clusters. Figure 4.2 shows the gene latent representations. The top row of the figure visualizes the first two latent dimensions of each gene cluster. For the MPN patient, we can clearly see the batch effects with respect to their stimulation status, especially when we examine the gene clusters with higher rankings. This is expected, as these gene clusters are supposed to be the "most" differentially expressed. In the less significant clusters (the clusters with the smaller labels), none of the reduced dimensionality methods suggest there may be more than one cluster and no batch effects with respect to the stimulation status is observed. We also note that the differences between the three dimensionality reduction methods are more striking between the PCA aprroach versus the other two.

In Section 4.3.3, we will also examine the marker genes for each of these identified cluster to determine what each group is. These groups may be trivial though, since we can clearly see that the two main clusters are due to the sample's TNF-α stimulation status.

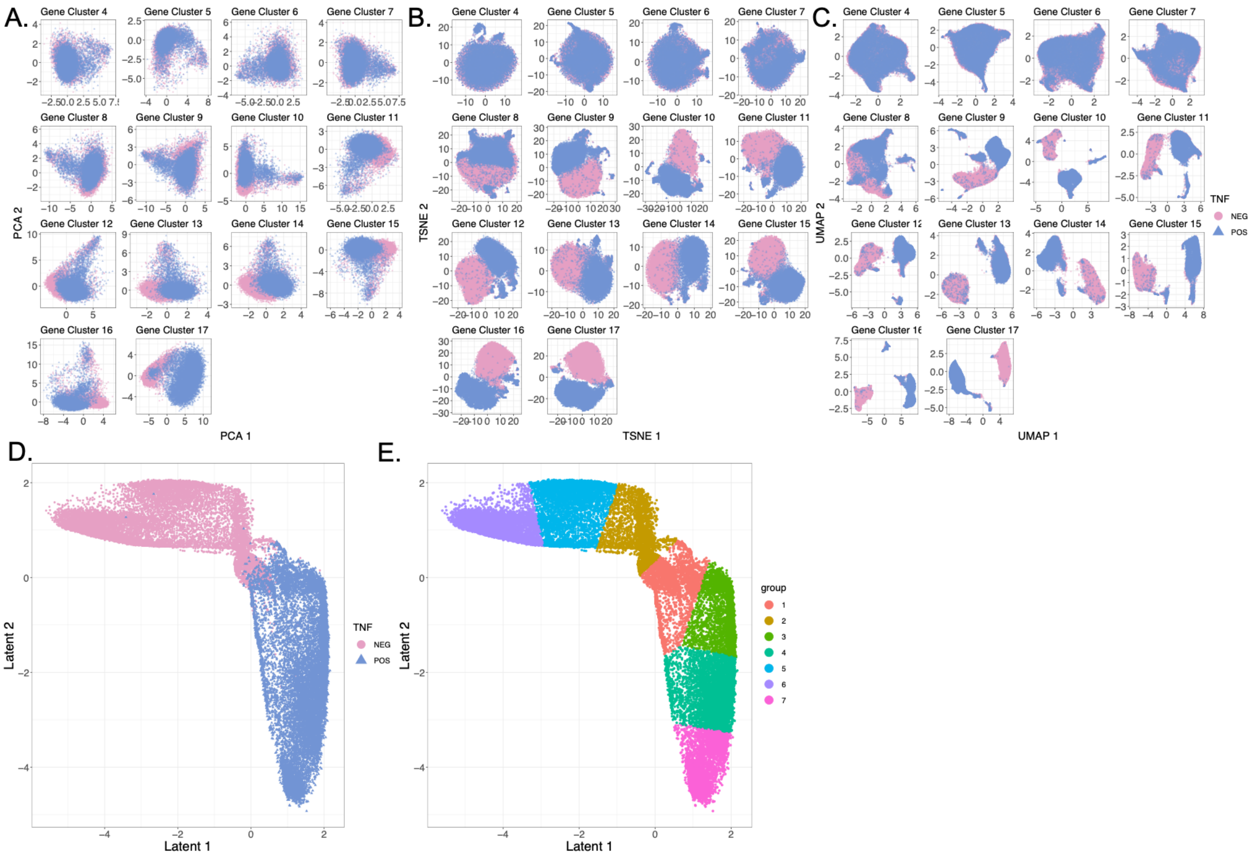


Figure 4.2: We visualize the latent representation of each gene cluster using PCA (A), *t*-SNE (B) and UMAP (C) for the MPN patient. After we obtain the gene latent representation, we can calculate the cell latent representation (D) and then cluster the cells (E).

4.3.3 Examining their marker genes

Figure 4.3 and Tables 4.1 and 4.2 show the top marker genes for each cluster in both patients. Only the genes with a positive average log2 fold change are shown (average log-normalized expression is higher in the cluster of interest relative to all other clusters). When we examine the heatmaps, it is clear that many of the clusters can be combined due to shared gene marker expression. However, the expression of Cluster 1 in the MPN patient (Figure 4.3B, suggests that cluster may be subdivided into two.

In the MPN patient, the marker genes in cluster 1 are enriched for GO terms *hydrogen peroxide catabolic process* (GO:0042744); *negative regulation of signal transduction in absence of ligand* (GO:1901099); and *negative regulation of extrinsic apoptotic signaling pathway in absence of ligand* (GO:2001240). These genes are also associated with signaling pathways apoptosis (KEGG 2021 Human) and the p53 signaling pathway (BioPlanet 2019). Cluster 1 is predominantly TNF-α stimulated cells. The other TNF-α stimulated cell clusters (Clusters 3, 4, and 7) are also enriched for GO terms *cytokine-mediated signaling pathway* (GO:0019221); *regulation of apoptotic process* (GO:0042981) and *cellular response to tumor necrotic factor* (GO:0071356). These clusters are significantly associated with the NF-κB signaling and apoptosis pathways (KEGG 2021 Human) and TNF-α effects on cytokine activity, cell motility and apoptosis (BioPlanet 2019). Cluster 7 is also enriched for *chemokine-mediated signaling pathway* (GO:0070098) and *cellular response to chemokine* (GO:1990869), and the chemokine signaling pathway (KEGG 2021 Human).

For the unstimulated cells, the marker genes for cluster 2 are enriched for *regulation of stem cell differentiation* (GO:2000736) and associated with the transcriptional misregulation in cancer (KEGG 2021 Human) and EGFR1 pathways (BioPlanet 2019). The marker genes for Cluster 5 are enriched for *cellular response to tumor necrosis factor* (GO:0071356) and upregulated in the GABAergic synapse (KEGG 2021 Human) and EGFR1 (BioPlanet 2019) pathways. Finally, the markers for Cluster 6 are enriched for *the ERK1 and ERK2 cascade* (GO:0070371) and upregulated in the transcriptional misregulation in cancer (KEGG 2021 Human) and vasopressin-like receptors and eIF4E release (BioPlanet 2019) pathways.

In the unaffected patient, there are no distinct unstimulated and stimulated clusters. The top enriched GO term for each cluster are: (1) *nucleotide-binding domain, leucine rich repeat containing receptor signaling pathway* (GO:0035872); (2) *cytokine-mediated signaling pathway* (GO:0019221); (3) *definitive hemopoiesis* (GO:0060216); (4) *regulation of apoptotic process* (GO:0042981) and *regulation of cell population proliferation* (GO:0042127); (5) *cytokine-mediated signaling pathway* (GO:0019221); (6) *cellular response to cytokine stimulus* (GO:0071345); and (7) *cellular response to hormone stimulus* (GO:0032870).

Regarding cell types/stage, the majority of these markers are not canonical human stem cell markers. We verified that HSC markers *CD34, THY1* and *ITGA6* as well as progenitor marker *PTPRC* were detected in both the unaffected and MPN samples, but these are not differentially expressed (as expected). There are some cell clusters that express known markers. For example, MPN cluster 1 appears to be erythocyte-like cells (*HBB, HBD*); MPN cluster 2 appears to be B-like cells (*LMO2*), cluster 5 appears to be endothelial-like cells (*CRHBP*), and cluster 6 progenitor cells (*EIF4EBP1, NME1*). In the unaffected patient, cluster 1 appears to be lymphoid-primed cells (*ZFP36L2*), cluster 4 appears to be T-like cells (*CCL4L2, JUNB, IFIT3*), and cluster 6 esinophil-like cells (*CXCL8*).

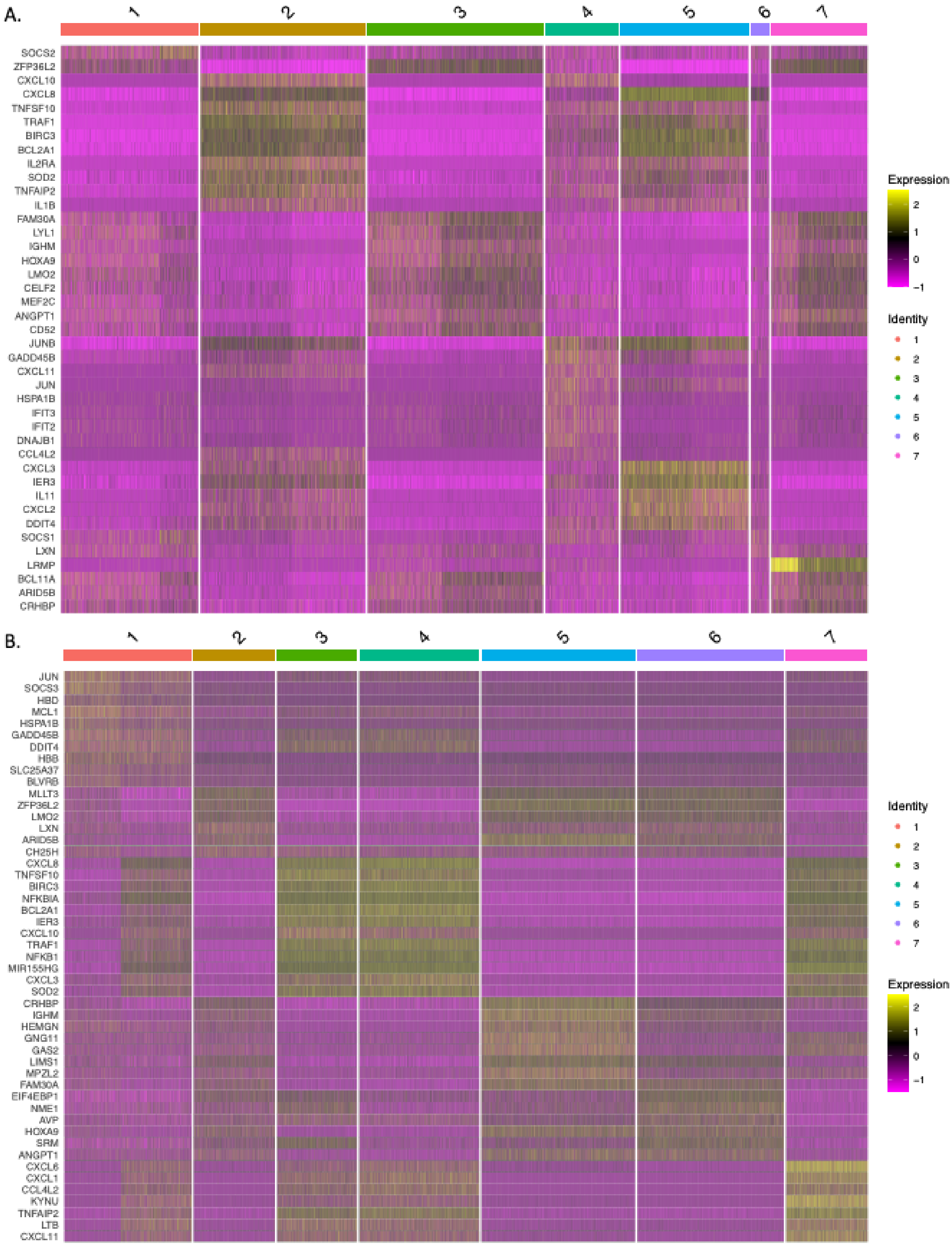


Figure 4.3: Heatmap of the top 10 marker genes for each cluster in the unaffected patient (A) and MPN patient (B).

Cluster	Gene	Adj p-value	Avg log2FC
1	*SOCS2*	6.32e-85	0.334
1	*ZFP36L2*	8.69e-73	0.274
2	*CXCL10*	0	0.953
2	*CXCL8*	0	0.939
2	*TNFSF10*	0	0.877
2	*TRAF1*	0	0.857
2	*BIRC3*	0	0.844
3	*ZFP36L2*	0	0.643
3	*FAM30A*	0	0.461
3	*LYL1*	5.52e-300	0.361
3	*HOXA9*	1.24e-294	0.373
3	*LMO2*	9.159e-288	0.391
4	*CXCL10*	2.39e-161	0.776
4	*GADD45B*	2.42e-111	0.617
4	*CXCL11*	7.8e-111	0.61
4	*JUN*	1.32e-107	0.572
4	*IFIT2*	1.13e-77	0.566
5	*CXCL8*	0	1.31
5	*CXCL3*	0	1.08
5	*BIRC3*	0	0.952
5	*IER3*	0	0.952
5	*BCL2A1*	0	0.849
6	*CXCL8*	1.05e-24	0.486
6	*SOCS1*	7.94e-09	0.251
6	*LXN*	7.2e-05	0.265
7	*LRMP*	0	0.705
7	*ZFP36L2*	0	0.584
7	*FAM30A*	4.94e-220	0.414
7	*ANGPT1*	2.29e-198	0.382
7	*CRHBP*	2.02e-152	0.366

Table 4.1: Top 5 marker genes for each cell cluster in the unaffected patient's samples. Only marker genes with a positive log2 fold change are shown. Note that some clusters may have less than five (positive) marker genes.

Cluster	Gene	Adj *p*-value	Avg log2FC
1	*JUN*	0	0.576
1	*SOCS3*	7.16e-296	0.568
1	*HBD*	5.68e-221	0.621
1	*GADD45B*	3.09e-122	0.422
1	*HBB*	1.07e-102	0.845
2	*MLLT3*	6.23e-285	0.337
2	*ZFP36L2*	6.46e-226	0.375
2	*LMO2*	3.89e-176	0.262
2	*LXN*	7.89e-160	0.318
2	*CH25H*	2.33e-68	0.282
3	*CXCL8*	0	0.913
3	*TNFSF10*	0	0.827
3	*BIRC3*	0	0.76
3	*NFKBIA*	0	0.749
3	*BCL2A1*	0	0.732
4	*CXCL8*	0	1.11
4	*BIRC3*	0	0.904
4	*NFKBIA*	0	0.858
4	*CXCL3*	0	0.837
4	*BCL2A1*	0	0.831
5	*CRHBP*	0	0.638
5	*ZFP36L2*	0	0.635
5	*IGHM*	0	0.586
5	*ARID5B*	0	0.543
5	*HEMGN*	0	0.522
6	*ZFP36L2*	0	0.493
6	*EIF4EBP1*	0	0.388
6	*NME1*	0	0.388
6	*AVP*	0	0.379
6	*MLLT3*	0	0.373
7	*CXCL6*	0	1.14
7	*CXCL1*	0	0.854
7	*CCL4L2*	0	0.795
7	*MIR155HG*	0	0.794
7	*KYNU*	0	0.794

Table 4.2: Top 5 marker genes for each cell cluster in the MPN patient's samples. Only marker genes with a positive log2 fold change are shown.

4.4 Discussion

In Chapter 2, we investigated the differential gene expression between an untreated sample and a sample treated with TNF-α in an unaffected patient and a MPN patient. We confirm that the MPN patient shows more response to inflammation than the unaffected patient. Here, we examine the response taking the heterogeneity of the cells into account.

Given the gene groupings from Chapter 2, we reduce the dimensionality of each gene group to one and concatenate their latent dimensions to form a cell latent representation. To identify which cells have a stronger relationship with each other given its gene latent space expression profile, we calculate the Pearson correlation matrix and cluster the cells.

Our framework is able to incorporate differentially expressed genes between two samples and more information from the genes overall to inform our cell clustering. Furthermore, as the amount of cells increase due to advances in sequencing technology, our framework can be updated to address computational limitations. For example, after obtaining the cell-cell correlation matrix, we can fit an autoencoder to the correlation matrix and extract the middle encoded layer. With an autoencoder, we can minimize the reconstruction loss between the original object (the cell-cell correlation matrix) and the decoded object (the correlation matrix projected from the middle encoded layer). This gives us higher confidence that our encoded latent space is a good representation of the original correlation matrix. However, with smaller correlation matrices such as the one for the unaffected patient, the clustering results are comparable when using UMAP compared to an autoencoder. So we decided to use UMAP to reduce the dimensionality of the correlation matrix for computational efficiency.

We are also currently using a fairly basic approach to cluster the reduced correlation matrix. Other methods, such as the Dirichlet process mixture model, may be more suited to determine the "correct" number of clusters given the data. The Dirichlet process mixture model is computationally expensive and has invariance issues when the input data is large, but the reduced correlation matrix has a dimension of $n_{cells} \times 2$, making it a good potential candidate for non-parametric clustering techniques.

As this “biclustering” framework has a nested approach, one major limitation is the gene clusters. In the previous chapter, we described a computational framework to obtain gene clusters given that we wanted to compare two samples. The resulting gene clusters are ranked groups of differentially expressed genes, with varying biological processes. If we had only one sample to “bicluster”, a natural set of gene groupings might be leveraging known pathway information to group the genes.

Chapter 5

Pipeline to predict the circadian time from genomics data

Junyan Duan[1], Michelle N. Ngo[1], Satya Swaroop Karri, Babak Shahbaba, John Lowengrub, Bogi Andersen

[1] Equal contribution

5.1 Introduction

Organisms have evolved intrinsic circadian clocks that help them anticipate and adjust to environmental changes caused by the 24-hour rotation of the Earth [7, 20]. The mammalian circadian clock is a biochemical oscillator powered by transcription-translation loops consisting of a positive arm and a negative arm [7, 20, 112]. In the positive arm, BMAL1 and CLOCK promote the expression of clock-controlled genes including the negative arm factors PER and CRY. PER and CRY inhibit the activating effect of BMAL1-CLOCK, leading to

24-hour oscillations.

In mammals, the suprachiasmatic nucleus (SCN) of the hypothalamus acts as the central pacemaker that coordinates and synchronizes circadian rhythms in peripheral tissues through neuronal and hormonal signals [14]. Besides signals from the SCN, environmental signals such as temperature [14], feeding [86, 124], and direct light [13] can selectively set peripheral clocks, sometimes causing asynchrony between the central and peripheral clocks. Epidemiological studies of shift workers and chronically jet-lagged individuals have revealed correlations between circadian disruption and cardiovascular diseases [82], mental health disorders [123], metabolic diseases [61, 89, 120], as well as cancer in various organs [75, 104] including skin [44, 134], breast [101, 45], and prostate [26, 128].

The goal of the nascent field of circadian medicine is to take into account circadian rhythm and its disruption in patient care. As the rhythm of a patient or diseased tissue is not necessarily synchronized with the external light-dark cycle, an important challenge in circadian medicine is to determine the internal circadian time of the patient or the tissue of focus. Such information can determine optimal time of treatment and identify conditions that might benefit from restoring circadian function [110, 28]. Current methods of circadian rhythm determination for a patient include the dim-light melatonin-onset assay [60], as well as circadian rhythm inference from body temperature [96] or cortisol levels in sweat [117].

Several groups have developed methods to infer circadian time of a sample (organism, organ, or tissue) based on transcriptomic data. ZeitZeiger [55] identifies useful features (genes) for prediction, scales the features over time, applies sparse principal component analysis, and predicts according to maximum likelihood estimation. BIO_CLOCK [1] uses supervised deep neural networks with coupled sine and cosine output units. TimeSignatR [10] is mainly designed for blood samples and applies within-subject renormalization and an elastic net predictor, making it generalizable between transcriptomic data from different assay platforms.

All existing methods have limitations. ZeitZeiger frequently runs into linear dependency issues, needs to be retrained before each prediction, and is not generalizable between transcriptomic platforms. BIO_CLOCK does not require re-training for each prediction but is not time-efficient. TimeSignatR performs well if there are two test samples, but performance depends on the time interval between the two samples. When given only one test sample, TimeSignatR can infer a second sample, but it has very low accuracy.

To address these issues, we present tauFisher, a pipeline that can accurately predict circadian time from a single transcriptomic data irrespective of the transcriptomic platform. tauFisher improves on previous methods in several ways: (1) it does not require the training data to be a complete time series; (2) the within-sample normalization step allows tauFisher to give an accurate prediction from just one sample; (3) since tauFisher only needs a few features to make accurate predictions, training and testing are computationally efficient; (4) tauFisher is platform agnostic and users only need to train the predictor once and can use the same predictor to make predictions for external data sets of the same tissue, regardless of the assay methods; and (5) unprecedentedly, tauFisher trained on bulk sequencing data is able to accurately predict the circadian time of single cell RNA sequencing (scRNA-seq) data.

To benchmark tauFisher, we begin by using tauFisher to predict the circadian time of eight previously studied data sets from multiple tissue types and experimental settings. We create several simulated time series data sets to demonstrate that tauFisher is not only able to accurately predict the circadian time of a sample but can also be used to investigate circadian phase heterogeneity in different cell types. We then collected a time series of scRNA-seq data from mouse dermis in this study. We found that most of the rhythmic processes are metabolism-related in dermal fibroblasts, while almost all rhythmic processes are related to immune response in dermal immune cells. Additionally, we found that the amplitude of the collective rhythm is dampened in dermal immune cells compared to dermal fibroblasts. Incorporating tauFisher with bootstrapping revealed that circadian phase heterogeneity contributes to the dampened collective rhythm as well as fewer rhythmic genes found in dermal immune cells.

5.2 Results

5.2.1 Overview of tauFisher

tauFisher is an assay platform-agnostic method that predicts circadian time from a single transcriptomic sample. The training part of the pipeline requires a time series of transcriptomic data and consists of five main steps: (1) identifying diurnal genes with a period length of 24 hours, (2) curve fitting using functional data analysis to fill in the missing time points and to make the training data less noisy, (3) within-sample normalization by calculating and scaling the difference in expression for each pair of predictor genes, (4) linearly transforming the scaled differences using principal component analysis, and (5) fitting a multinomial regression on the first two principal components (Figure 5.1, Method Section 5.3.1).

For testing, a transcriptomic sample without a time label is narrowed to include only the predictor genes identified in the training data. After the within-sample normalization step, the test sample is projected to the principal component space and multinomial regression is performed to predict time of the test sample (Figure 5.1, Method Section 5.3.1).

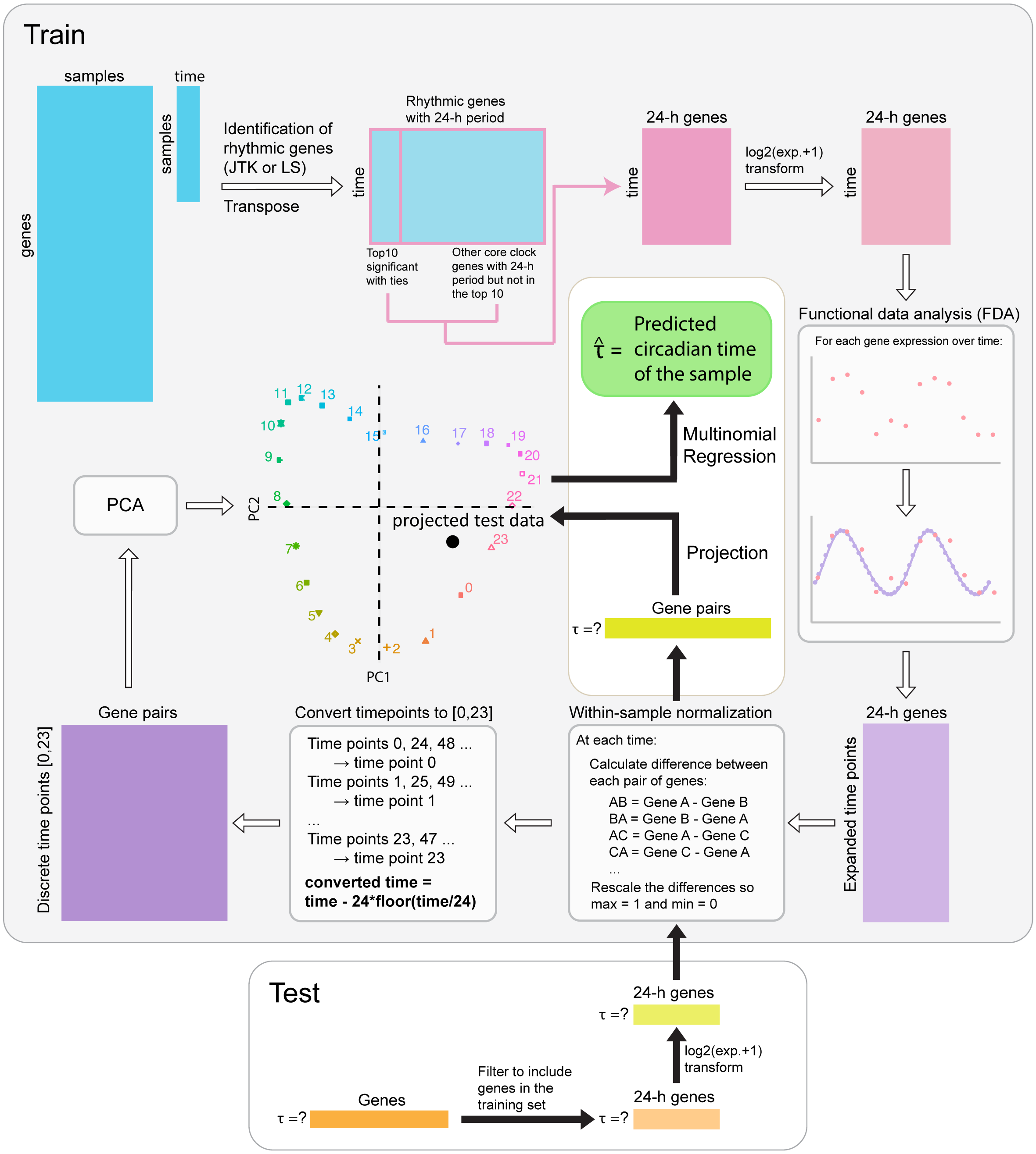

Train
samples
time
genes
samples
Identification of rhythmic genes (JTK or LS)
Transpose
Rhythmic genes with 24-h period
time
Top10 significant with ties
Other core clock genes with 24-h period but not in the top 10
24-h genes
time
log2(exp.+1) transform
24-h genes
time
Functional data analysis (FDA)
For each gene expression over time:
$\hat{\tau}$ = Predicted circadian time of the sample
Multinomial Regression
Projection
Gene pairs
τ =?
PCA
PC2
PC1
projected test data
11
12
13
14
10
15
16
17
18
19
9
20
21
8
22
23
7
6
0
5
1
4
3
2
Gene pairs
Discrete time points [0,23]
Convert timepoints to [0,23]
Time points 0, 24, 48 ...
→ time point 0
Time points 1, 25, 49 ...
→ time point 1
...
Time points 23, 47 ...
→ time point 23
converted time =
time - 24*floor(time/24)
Within-sample normalization
At each time:
Calculate difference between each pair of genes:
AB = Gene A - Gene B
BA = Gene B - Gene A
AC = Gene A - Gene C
CA = Gene C - Gene A
...
Rescale the differences so max = 1 and min = 0
24-h genes
Expanded time points
Test
24-h genes
τ =?
log2(exp.+1) transform
Genes
τ =?
Filter to include genes in the training set
τ =?
24-h genes

Figure 5.1: Key steps of the tauFisher pipeline given an expression matrix ("exp.") includes identification of periodic genes, functional data analysis, within-sample feature normalization, and multinomial regression.

5.2.2 tauFisher outperforms current methods when trained and tested on bulk-sequencing data.

To assess the robustness and accuracy of tauFisher in predicting circadian time from a single sample of transcriptomic data, we applied tauFisher to a diverse set of data collected from different species, tissues and sequencing platforms (Table 5.1).

Table 5.1: data sets from different species, tissues and sequencing platforms were used to benchmark tauFisher's ability to predict circadian time.

Data	**Year**	**GEO**	**Species**	**Tissue**	**Platform**	**Sampling Frequency**	**Time Course Duration**
Arnardottir ES *et al.* [6]	2014	GSE56931	*Homo sapiens*	Blood	Custom Affymetrix Microarray	4h	72h
Braun R *et al.* [10]	2018	GSE113883	*Homo sapiens*	Blood	Illumina NextSeq 500	2h	28h
Geyfman M *et al.* [39]	2012	GSE38622	*Mus musculus*	Skin	Affymetrix Mouse Gene 1.0 ST Array	4h	48h
Tognini P *et al.* [115]	2020	GSE157077	*Mus musculus*	SCN	Illumina HiSeq 4000	4h	24h
Zhang R *et al.* [137]	2014	GSE54650	*Mus musculus*	Kidney	Affymetrix Mouse Gene 1.0 ST Array	2h	48h
Zhang R *et al.* [137]	2014	GSE54650	*Mus musculus*	Liver	Affymetrix Mouse Gene 1.0 ST Array	2h	48h
Zhang R *et al.* [137]	2014	GSE54651	*Mus musculus*	Kidney	Illumina HiSeq 2000	6h	48h
Zhang R *et al.* [137]	2014	GSE54651	*Mus musculus*	Liver	Illumina HiSeq 2000	6h	48h

For each benchmark data set, we generated 100 random train and test partitions (without replacement) of the samples. In each partition, we used 80% of the samples for training and 20% for testing.

We compared tauFisher to the current state-of-the-art methods: ZeitZeiger [55] and TimeSignatR [10]. As TimeSignatR is able to predict circadian time for samples from one time point alone, or one time point with an inferred second time point, we included both types of predictions in the benchmark.

We define a prediction within two hours of the true time to be correct. Using other time ranges to define correctness minimally change the benchmark outcome (Table 5.3).

tauFisher achieved the highest accuracy for all eight benchmark data sets; seven using predictor genes found by JTK_Cycle [54] and one using Lomb-Scargle [41] (Figure 5.2; Table 5.2). While ZeitZeiger did not attain the highest accuracy in any of the data sets, it achieved the lowest root mean squared error (RMSE) in three of the data sets and comparable RMSE to tauFisher in half of the data sets. However, ZeitZeiger could not predict the time for several iterations due to linearly dependent basis vectors. Particularly, in the kidney and liver bulk sequencing data sets, ZeitZeiger failed to predict the time for all 100 iterations. TimeSignatR performed the worst among the three methods, giving the lowest accuracy and the highest RMSE.

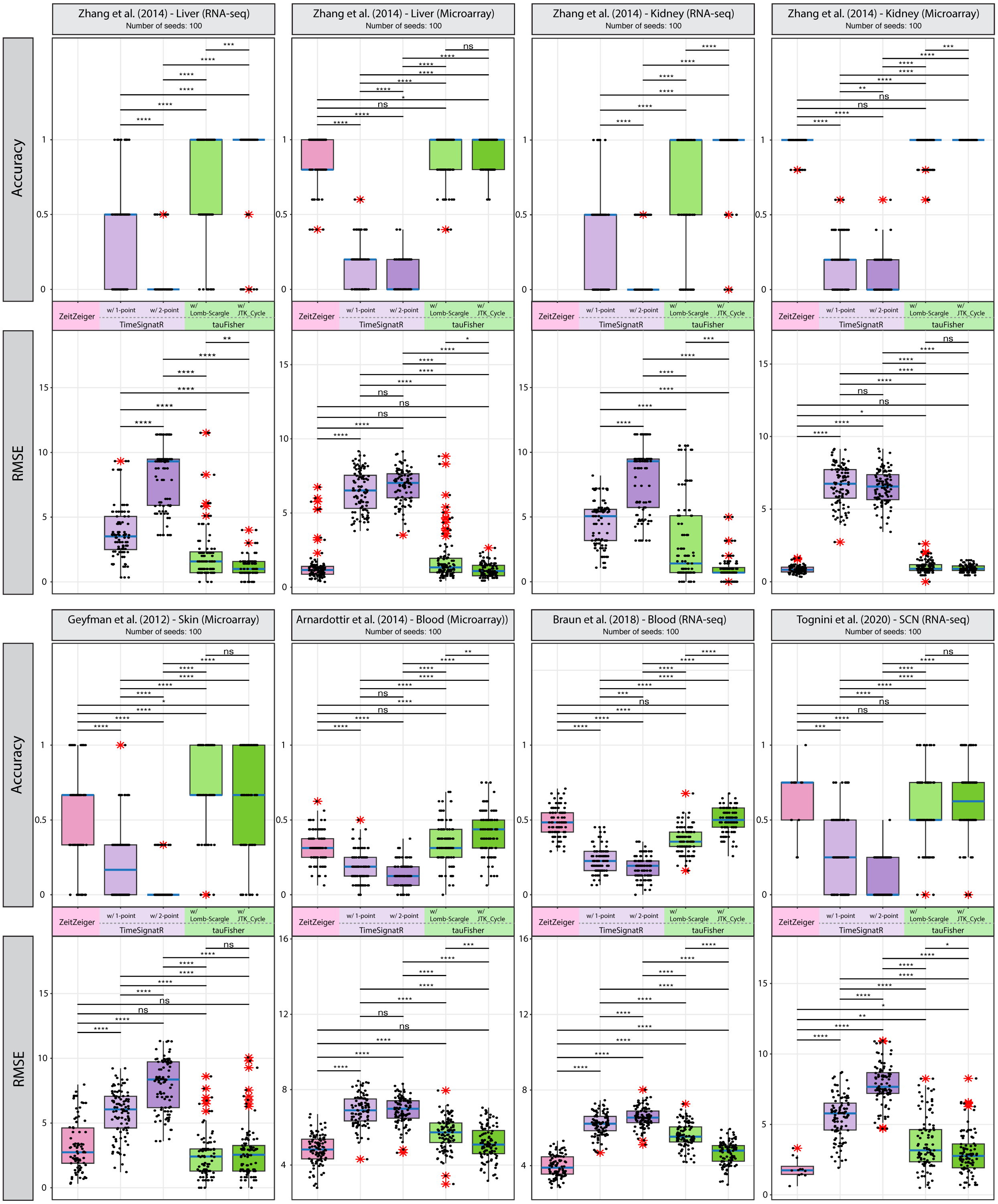


Figure 5.2: tauFisher outperforms published circadian time prediction methods in both accuracy and RMSE for transcriptomic data collected from various organs and assay platforms. ns: p-value > 0.05, $*$: p-value ≤ 0.05, p-value ≤ 0.01, $***$: p-value ≤ 0.001, $****$: p-value ≤ 0.0001.

Table 5.2: Benchmark results (mean ± standard deviation) when we train on 80% of the data set and predict the circadian time of 20%.

Data	Metric	ZeitZeiger	TimeSignatR		tauFisher	
			w/ 1-point	w/ 2-point	w/ Lomb-Scargle	w/ JTK_Cycle
[6]	Accuracy	0.325 ± 0.103	0.168 ± 0.110	0.126 ± 0.084	0.347 ± 0.139	**0.417 ± 0.148**
	RMSE	**4.829 ± 0.779**	7.213 ± 1.440	7.209 ± 1.403	5.666 ± 0.848	5.167 ± 0.890
	# NA	0	6	6	0	0
[10]	Accuracy	0.487 ± 0.094	0.235 ± 0.087	0.181 ± 0.077	0.367 ± 0.088	**0.499 ± 0.084**
	RMSE	**3.957 ± 0.571**	6.198 ± 0.574	6.572 ± 0.517	5.606 ± 0.646	4.655 ± 0.610
	# NA	0	0	0	0	0
[40]	Accuracy	0.460 ± 0.344	0.210 ± 0.240	0.047 ± 0.116	**0.740 ± 0.258**	0.670 ± 0.330
	RMSE	4.558 ± 3.640	5.781 ± 1.785	7.982 ± 1.949	**2.488 ± 1.662**	2.707 ± 1.976
	# NA	15	0	0	0	0
[115]	Accuracy	0.075 ± 0.218	0.285 ± 0.228	0.075 ± 0.120	0.545 ± 0.239	**0.635 ± 0.237**
	RMSE	10.784 ± 3.317	5.540 ± 1.512	7.883 ± 1.392	3.722 ± 1.880	**2.969 ± 1.477**
	# NA	88	0	0	0	0
[137][2,4]	Accuracy	0.986 ± 0.051	0.176 ± 0.159	0.100 ± 0.119	0.966 ± 0.086	**1.000 ± 0.000**
	RMSE	**0.849 ± 0.244**	6.639 ± 1.366	6.511 ± 1.172	1.021 ± 0.406	0.911 ± 0.256
	# NA	0	0	0	0	0
[137][1,2,5]	Accuracy	`NA`	0.330 ± 0.286	0.100 ± 0.201	0.695 ± 0.414	**0.945 ± 0.200**
	RMSE	`NA`	4.578 ± 1.733	8.136 ± 2.638	2.965 ± 3.117	**1.076 ± 0.981**
	# NA	100	0	0	0	0
[137][3,4]	Accuracy	0.868 ± 0.156	0.164 ± 0.137	0.076 ± 0.106	0.880 ± 0.161	**0.932 ± 0.103**
	RMSE	1.391 ± 1.124	6.467 ± 1.382	6.844 ± 1.137	1.797 ± 1.479	**1.182 ± 0.451**
	# NA	0	0	0	0	0
[137][1,3,5]	Accuracy	`NA`	0.360 ± 0.341	0.065 ± 0.169	0.765 ± 0.344	**0.910 ± 0.269**
	RMSE	`NA`	3.787 ± 2.067	8.343 ± 2.434	2.269 ± 2.356	**1.203 ± 0.825**
	# NA	100	0	0	0	0

[1] If ZeitZeiger [55] was unable to do a 3-fold cross validation, we ran ZeitZeiger without any cross validation and set `sumabsv` = 1 and `nSpc` = 3. [2] kidney [3] liver [4] microarray [5] bulk RNA

Table 5.3: Benchmark results (mean ± standard deviation) when we train on 80% of the data set and predict the circadian time of 20%.

Data	Metric	ZeitZeiger	TimeSignatR		tauFisher	
			w/ 1-point	w/ 2-point	w/ Lomb-Scargle	w/ JTK_Cycle
[6]	Accuracy (within 3 hr)	0.480 ± 0.122	0.250 ± 0.131	0.181 ± 0.098	0.456 ± 0.151	**0.517 ± 0.135**
	Accuracy (within 2 hr)	0.325 ± 0.103	0.168 ± 0.110	0.126 ± 0.084	0.347 ± 0.139	**0.417 ± 0.148**
	Accuracy (within 1 hr)	0.174 ± 0.094	0.085 ± 0.095	0.077 ± 0.076	0.209 ± 0.107	**0.264 ± 0.129**
	Accuracy (exact)	0.002 ± 0.011	0.000 ± 0.000	0.000 ± 0.000	0.076 ± 0.075	**0.104 ± 0.078**
[10]	Accuracy (within 3 hr)	**0.635 ± 0.087**	0.342 ± 0.097	0.266 ± 0.085	0.477 ± 0.096	0.625 ± 0.078
	Accuracy (within 2 hr)	0.487 ± 0.094	0.235 ± 0.087	0.181 ± 0.077	0.367 ± 0.088	**0.499 ± 0.084**
	Accuracy (within 1 hr)	0.250 ± 0.067	0.120 ± 0.061	0.084 ± 0.053	0.235 ± 0.073	**0.304 ± 0.071**
	Accuracy (exact)	0.001 ± 0.006	0.000 ± 0.000	0.000 ± 0.000	0.072 ± 0.053	**0.092 ± 0.052**
[39]	Accuracy (within 3 hr)	0.563 ± 0.350	0.320 ± 0.259	0.070 ± 0.136	**0.783 ± 0.248**	0.750 ± 0.286
	Accuracy (within 2 hr)	0.460 ± 0.344	0.210 ± 0.240	0.047 ± 0.116	**0.740 ± 0.258**	0.670 ± 0.330
	Accuracy (within 1 hr)	0.307 ± 0.275	0.107 ± 0.163	0.027 ± 0.091	**0.590 ± 0.280**	0.543 ± 0.295
	Accuracy (exact)	0.010 ± 0.057	0.000 ± 0.000	0.000 ± 0.000	**0.257 ± 0.263**	0.227 ± 0.241
[115]	Accuracy (within 3 hr)	0.108 ± 0.298	0.402 ± 0.230	0.100 ± 0.133	0.610 ± 0.237	**0.705 ± 0.237**
	Accuracy (within 2 hr)	0.075 ± 0.218	0.285 ± 0.228	0.075 ± 0.120	0.545 ± 0.239	**0.635 ± 0.237**
	Accuracy (within 1 hr)	0.040 ± 0.150	0.128 ± 0.172	0.032 ± 0.084	0.412 ± 0.257	**0.468 ± 0.253**
	Accuracy (exact)	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	**0.190 ± 0.195**	0.182 ± 0.181
[137][2,4]	Accuracy (within 3 hr)	0.996 ± 0.028	0.256 ± 0.178	0.192 ± 0.133	0.996 ± 0.028	**1.000 ± 0.000**
	Accuracy (within 2 hr)	0.986 ± 0.051	0.176 ± 0.159	0.100 ± 0.119	0.966 ± 0.086	**1.000 ± 0.000**
	Accuracy (within 1 hr)	0.728 ± 0.179	0.096 ± 0.125	0.012 ± 0.048	0.880 ± 0.158	**0.916 ± 0.128**
	Accuracy (exact)	0.004 ± 0.028	0.000 ± 0.000	0.000 ± 0.000	**0.370 ± 0.219**	0.358 ± 0.209
[137][1,2,5]	Accuracy (within 3 hr)	NA	0.475 ± 0.279	0.115 ± 0.211	0.720 ± 0.416	**0.945 ± 0.200**
	Accuracy (within 2 hr)	NA	0.330 ± 0.286	0.100 ± 0.201	0.695 ± 0.414	**0.945 ± 0.200**
	Accuracy (within 1 hr)	NA	0.240 ± 0.251	0.070 ± 0.174	0.565 ± 0.453	**0.815 ± 0.346**
	Accuracy (exact)	NA	0.000 ± 0.000	0.000 ± 0.000	0.250 ± 0.314	**0.375 ± 0.344**
[137][3,4]	Accuracy (within 3 hr)	0.964 ± 0.100	0.264 ± 0.186	0.168 ± 0.105	0.936 ± 0.133	**0.986 ± 0.059**
	Accuracy (within 2 hr)	0.868 ± 0.156	0.164 ± 0.137	0.076 ± 0.106	0.880 ± 0.161	**0.932 ± 0.103**
	Accuracy (within 1 hr)	0.626 ± 0.225	0.086 ± 0.118	0.000 ± 0.000	0.764 ± 0.196	**0.850 ± 0.162**
	Accuracy (exact)	0.002 ± 0.020	0.000 ± 0.000	0.000 ± 0.000	**0.312 ± 0.189**	0.308 ± 0.219
[137][1,3,5]	Accuracy (within 3 hr)	NA	0.465 ± 0.357	0.075 ± 0.179	0.840 ± 0.332	**0.970 ± 0.171**
	Accuracy (within 2 hr)	NA	0.360 ± 0.341	0.065 ± 0.169	0.765 ± 0.344	**0.910 ± 0.269**
	Accuracy (within 1 hr)	NA	0.240 ± 0.289	0.060 ± 0.163	0.585 ± 0.427	**0.735 ± 0.313**
	Accuracy (exact)	NA	0.000 ± 0.000	0.000 ± 0.000	0.155 ± 0.263	**0.335 ± 0.326**

[1] If ZeitZeiger [55] was unable to do a 3-fold cross validation, we ran ZeitZeiger without any cross validation and set `sumabsv` = 1 and `nSpc` = 3. [2] kidney [3] liver [4] microarray [5] bulk RNA

5.2.3 tauFisher accurately predicts circadian time for cross-platform bulk transcriptomic data.

Since tauFisher gives accurate circadian time prediction for bulk transcriptomic data collected from various platforms, we examined its performance when trained and tested on data sets generated from different platforms. We used rhythmic genes identified by JTK_Cycle in the tauFisher pipeline, since this combination resulted in the most accurate predictions in the within-platform benchmark.

We trained tauFisher on GSE38622 [39], a microarray data set collected from mouse dorsal skin every four hours for 48 hours under regular 12:12 light-dark cycle (zeitgeber time [ZT] 2, 6, 10, ...). The test data set is from GSE83855 [124], a bulk RNA-seq data set collected every four hours for 28 hours under 12:12 light-dark cycle (ZT0, 4, 8, ...) from mouse dorsal skin in a time-restricted feeding study. Since time of feeding influences tissue's circadian clock [124, 28], we only selected the ad libidum control condition for testing so that the time labels best represent the internal time.

Eighteen genes were selected to be predictor features. Though the train and test data sets are not on the same scale and were collected at different time points, their overall rhythmic patterns agree with each other (Figure 5.3A). For seven out of the eight tests, tauFisher predicted a circadian time that is within the 2-hour range from the actual time label, giving a high accuracy of 0.875 and a low RMSE of 2.669 (Figure 5.3A). This example demonstrates tauFisher's ability to accurately predict circadian time across bulk transcriptomics platforms.

5.2.4 tauFisher trained on bulk RNA-seq data and microarray data accurately predicts circadian time of scRNA-seq samples.

tauFisher's ability to predict circadian time is not limited to cross-platform bulk-level transcriptomic data sets. It can predict circadian time for scRNA-seq data as well. In particular, tauFisher only needs to be trained on a time series of bulk-level transcriptomic data, which is more abundant and cheaper to collect than a scRNA-seq data time series.

Since most published scRNA-seq data sets do not have time labels, the selection of test data sets was limited. Here we tested tauFisher on scRNA-seq data collected from the mouse SCN [127] and mouse dermal skin (collected in this study).

GSE117295 [127] includes twelve SCN samples collected from circadian time (CT) 14 to 58 every four hours (CT14, 18, 22, ...) under constant darkness, and one light-stimulated SCN sample. Since light immediately leads to differential expression of rhythmic genes [127], only the samples from the control experiment were used for benchmarking. For each of the twelve samples, a pseudobulk data set was generated for testing. For training, we chose GSE157077 [115], a time series of bulk RNA-seq data collected from the mouse SCN every four hours under regular 12:12 light-dark cycle starting at ZT0. Since each time point in the training data set contains three replicates, instead of averaging them, we concatenated the replicates so that the input training data spans 72 hours.

Twenty genes from the training data passed the feature selection criteria. These genes display robust rhythms in both the training data and the testing pseudobulk data (Figure 5.3B). The test data appears to be noisier since it is not normalized. tauFisher does not require the test data to be normalized as the within-sample normalization step is part of the pipeline.

In ten out of the twelve tests, tauFisher predicted a time that is within 2-hour of the labeled time, resulting a high accuracy of 0.833 and a low RMSE of 1.936 (Figure 5.3B). Although none of the published circadian time prediction methods claims to be able to predict time for scRNA-seq data, we still tested their performance. tauFisher greatly outperforms ZeitZeiger and TimeSignatR in both accuracy and RMSE (Table 5.4).

Table 5.4: tauFisher trained on SCN bulk RNA-seq data [115] accurately predicts circadian time for pseudobulk data (generated from scRNA-seq data [127]).

		ZeitZeiger		**TimeSignatR**				**tauFisher**	
				w/ 1pt		w/ 2pt		w/ JTK_Cycle	
Sample	Truth	Prediction	Error	Prediction	Error	Prediction	Error	Prediction	Error
CT14	14	16	-2	14	0	14	0	13	1
CT18	18	2	-8	12	6	15	3	17	1
CT22	22	2	-4	12	10	14	8	0	-2
CT26	2	2	0	11	-9	13	-11	0	2
CT30	6	15	-9	11	-5	12	-6	10	-4
CT34	10	15	-5	11	-1	12	-2	10	0
CT38	14	16	-2	14	0	14	0	13	1
CT42	18	2	-8	13	5	15	3	18	0
CT46	22	2	-4	13	9	16	6	0	-2
CT50	2	2	0	11	-9	16	10	0	2
CT54	6	15	-9	10	-4	11	-5	9	-3
CT58	10	15	-5	12	-2	12	-2	11	-1
Accuracy		0.33		0.33		0.33		**0.83**	
RMSE		5.63		6.12		5.83		**1.94**	

To ensure that tauFisher's performance on scRNA-seq data is consistent and applicable to peripheral clocks, we performed scRNA-seq on adult wild type C57BL/6J mouse dermis every four hours for 72 hours under 12:12 light-dark cycle for testing. The pseudobulk matrix for the 18 samples were computed directly from the unprocessed data. We trained tauFisher on GSE38622 [39], a time series of microarray data. Because two of the rhythmic genes, *A630005I04Rik* and *Ivl*, are not present in the pseudobulk data, only 16 features were selected in the tauFisher pipeline in this test (Figure 5.3C).

While the input test data, the unnormalized pseudobulk data, appears to be noisy, tauFisher successfully predict circadian times for the 18 samples thanks to the within-sample normalization step in the tauFisher pipeline (Figure 5.3C). In 14 out of the 18 tests, tauFisher predicted circadian time within 2 hours of the labeled time, giving a high accuracy of 0.778 and a low RMSE of 2.198 (Figure 5.3C).

In summary, we have demonstrated that tauFisher trained on bulk-level transcriptomic data, either bulk RNA-seq or microarray data, can accurately predict circadian time for scRNA-seq data sets, making it particularly useful for expanding the current scRNA-seq database for circadian studies by adding time labels to existing scRNA-seq data.

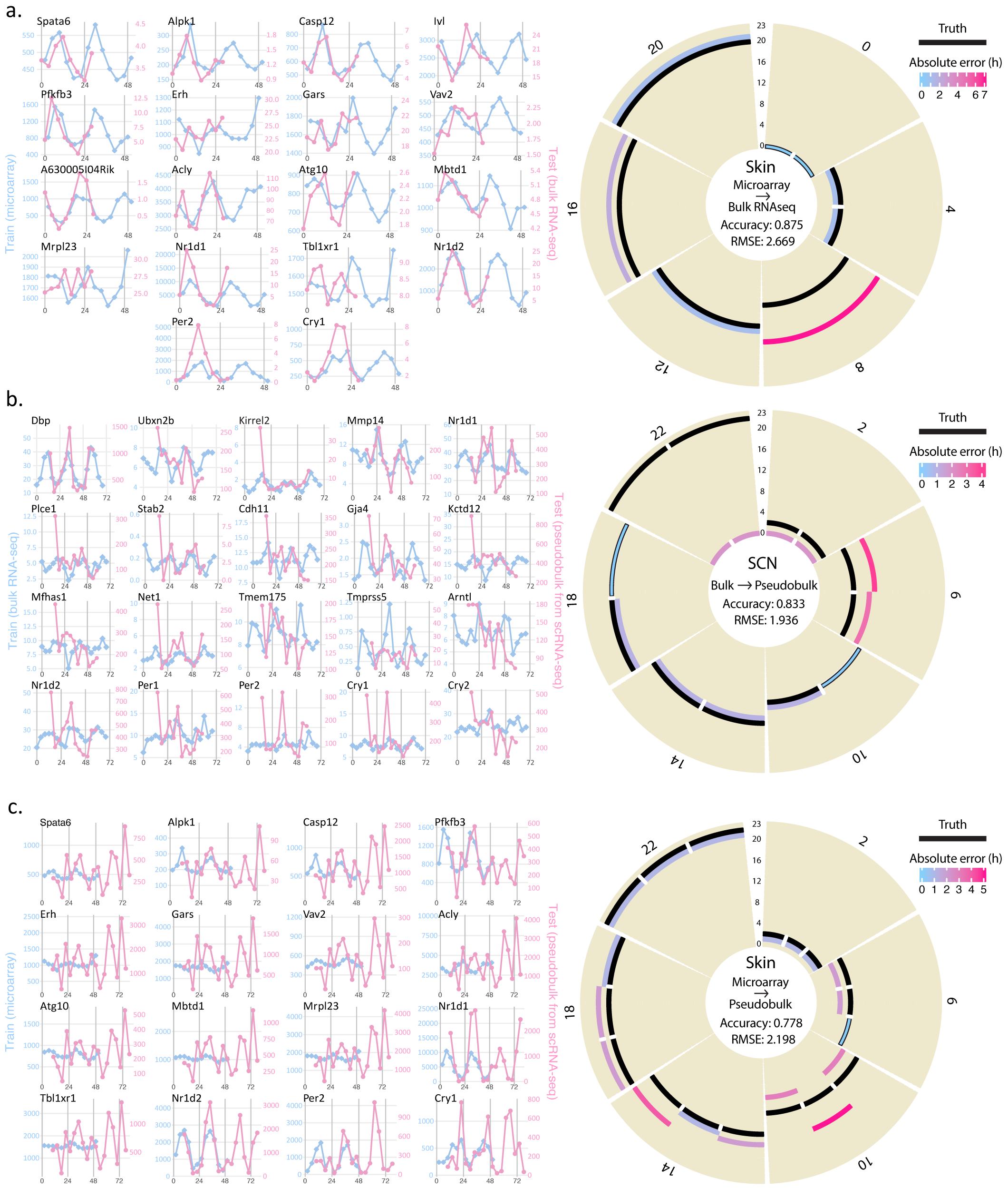


Figure 5.3: tauFisher accurately predicts circadian time even when the training and test data are from different assay methods. **a** tauFisher trained on skin microarray data can predict circadian time for skin bulk RNA-seq data. **b** tauFisher trained on SCN bulk RNA-seq data can predict circadian time of pseudobulk data generated from SCN scRNA-seq data. **c** tauFisher trained on skin microarray data can be used to predict circadian time of pseudobulk data generated from dermis scRNA-seq data. **a, b, c** Left: Predictor features and their expression in the training and test data. Right: Prediction outcomes for the tests. Labels on the circumference are the time labels for the test data. Predicted times are plotted along the radius, with colors representing absolute error. Black bars along the radius marks the truth (same as the labels on the circumference).

5.2.5 Collective circadian rhythms are dampened in dermal immune cells compared to dermal fibroblasts.

Due to the frequency of sequencing dropouts for clock genes in scRNA-seq data, investigating the circadian clock within each cell is not yet achievable. To overcome this limitation, previous studies have used pseudobulk approaches to investigate the clock in scRNA-seq data [127].

To validate the pseudobulk approach for studying the circadian clock in mouse dermis, we normalized the pseudobulk scRNA-seq data, and compared it with the published microarray data GSE38622 from mouse whole skin [39]. Overlay of the expression of nine core clock genes, *Arntl*, *Dbp*, *Per1*, *Per2*, *Per3*, *Nr1d1*, *Nr1d2*, *Cry1* and *Cry2*, reveals perfect consistency between the microarray data and the scRNA-seq data (Figure 5.4A), indicating that circadian clock gene expression in the dermis is captured in the pseudobulk data generated from scRNA-seq data.

To study the circadian clock at a cell-type level in the skin, we integrated all samples and performed scRNA-seq analysis to identify cell types. In total, 16,866 cells passed the quality control, with around 950 cells per sample and around 2,800 cells per ZT. Four major cell types, fibroblasts (N = 12,649), immune cells (N = 3,353), muscle cells (N = 722) and endothelial cells (N = 142) were identified using canonical marker genes (Figure 5.4B). Due to low cell counts for muscle and endothelial cells (N<20) in some samples, we could not generate a reliable time series of pseudobulk data for these two cell types. Thus, we focus on studying the circadian clock in dermal fibroblasts and immune cells in this study.

To compare the core clock in fibroblasts and immune cells, we computed and normalized the pseudobulk for each of the two cell types in each sample. Both fibroblasts and immune cells possess robust circadian clock at the pseudobulk-level. While the overall rhythms in the two cells types are consistent with each other, with core clock gene expressions peaking and troughing around the same time, the amplitudes of the oscillations are reduced in the immune cells compared to fibroblasts, indicating a dampened collective clock in immune cells (Figure 5.4C). Whether this observation indicates less synchronous clocks in immune cells than in fibroblasts, or weaker clock function in each individual immune cell, is not known.

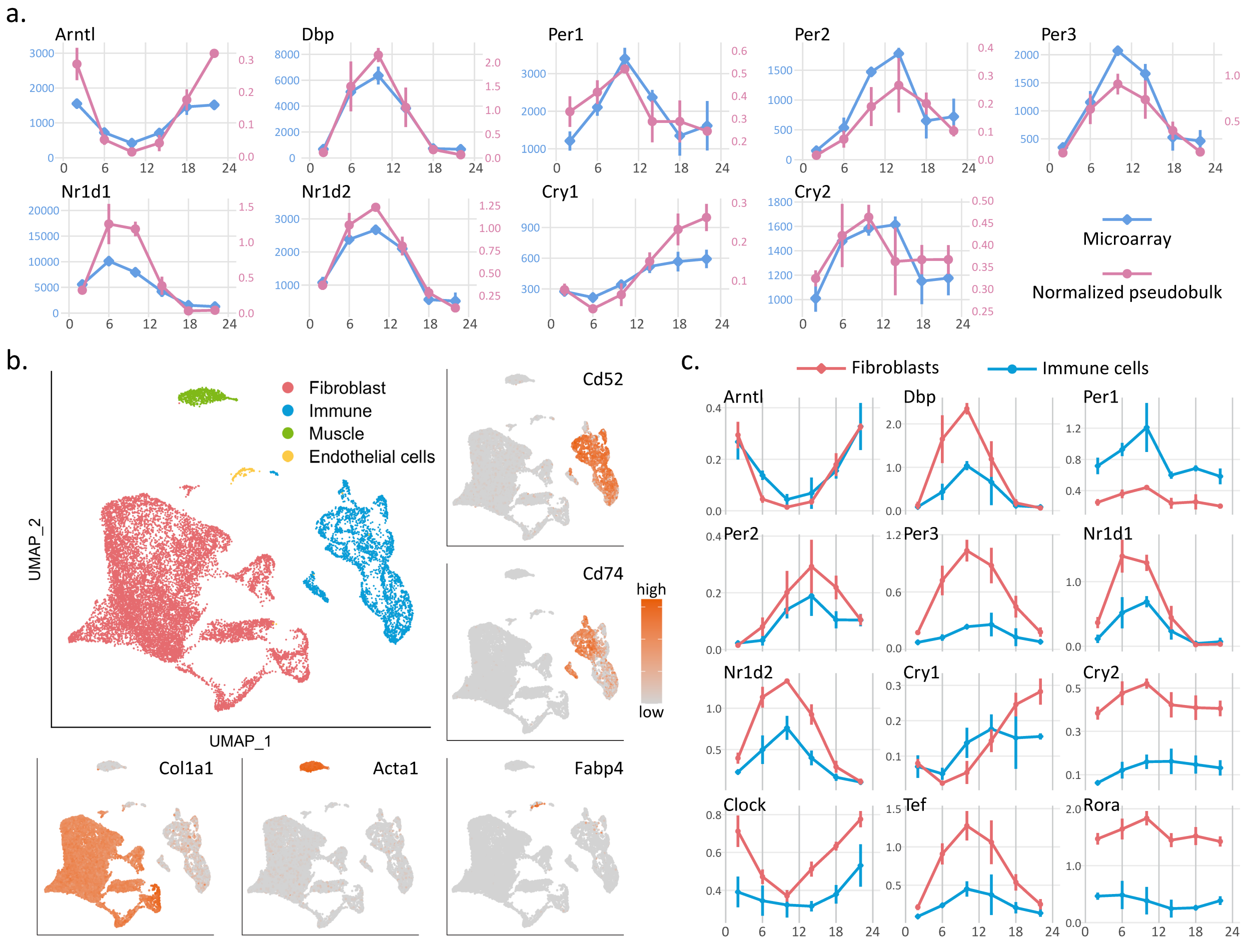


Figure 5.4: **a** Expression of the core clock genes in the normalized pseudobulk generated from scRNA-seq data (pink) is consistent with their expression in the published microarray data (blue). **b** Four major cell types, fibroblasts (red), immune cells (blue), muscle cells (green) and endothelial cells (yellow) are identified using canonical marker genes. Feature plots of the representative marker genes are shown (orange: high expression; grey: low expression); *Col1a1* for fibroblasts, *Acta1* for muscle cells, *Fabp4* for endothelial cells, *Cd52* and *Cd74* for immune cells. **c** At the pseudobulk-level, expression pattern of the core clock genes are similar in fibroblasts (red) and immune cells (blue), while the amplitudes of the oscillations are dampened in immune cells fore most of the core clock genes.

5.2.6 Dermal fibroblasts and immune cells harbor different rhythmic pathways and processes.

To study diurnal genes and pathways in dermal fibroblasts and immune cells, we used JTK_Cycle to identify rhythmic genes from the normalized pseudobulk data. We identified 1,867 and 353 rhythmic genes in fibroblasts and immune cells, respectively (Figure 5.5A). The fewer rhythmic genes in immune cells is not caused by the lower cell count of immune cells, as randomly down-sampling the fibroblasts to the number of immune cells produced similar results. Only 79 genes were rhythmic in both cell types, with most of them related to the core clock network and metabolism.

Gene Ontology analysis revealed that rhythmic processes in fibroblasts and immune cells are different. Shared terms reflect basic cell integrity maintenance and function including *nucleocytoplasmic transport*, *regulations of cellular amide metabolic process*, *regulation of protein stability*, and *rhythmic process* (Figure 5.5B). For fibroblasts, additional metabolism processes and migration are significantly enriched by the rhythmic genes (Figure 5.5B, red). For immune cells, the rhythmic genes enrich for more immune responses including *defense response to virus*, *regulation of T-helper 2 cell differentiation*, and *response to interferon-beta* (Figure 5.5B, blue).

We selected some of the rhythmic genes in fibroblasts (Figure 5.5C) and immune cells (Figure 5.5D)and compared their expression patterns in the two cell types. For fibroblasts, we highlight genes related to glucose metabolism (*Pkm*), glycosylation (*Gal3st4*, *Plpp3*), OXPHOS (*Ndufs8*), collagen (*Loxl2*, *Tgfb1*), amino acid metabolism (*Ivd*), sterol synthesis (*Scp2*, *Por*), and cell adhesion and migration (*Elmo2*, *Antxr1*), suggesting circadian regulation of the above processes at a molecular level (Figure 5.5C). Interestingly, while some genes are only significantly rhythmic in fibroblasts because they are not expressed in immune cells (e.g. *Loxl2*), some are expressed at similar or higher levels in immune cells, but are not significantly rhythmic in the latter (e.g. *Ndufs8*, *Scp2*), indicating cell-type specific circadian regulations.

For the immune cells, genes related to inflammatory and immune response (*Cdk19*, *Cd84*), post-translational modification (*Sumo1*), extracellular matrix regulation (*Mmp9*), transcription regulation (*Med16*), electrochemical gradient maintenance (*Atp1b1*), and intercellular communication (*Stxbp6*) are rhythmic (Figure 5.5D). We note that *Sumo1* is rhythmic in both fibroblasts and immune cells, but the expression peaks 4-hour later in immune cells than in fibroblasts.

Interestingly, the expression of *Il18r1* is significantly rhythmic with a large amplitude in fibroblasts (p-value = 2.21×10^{-7}), but not in immune cells (p-value = 0.7104) (Figure 5.5C). The level of IL18, the ligand that binds to IL18R1, was found to be rhythmic in mouse peripheral blood [67]. Here, *Il18*, the gene encodes IL18, is significantly rhythmic in neither fibroblasts (p-value = 0.3097) nor immune cells (p-value = 0.0925) (Figure 5.5D). But, it is possible that the insignificance of the p-value for immune cells is caused by noise introduced by summing the expression of all types of immune cells.

To further explore the rhythmic pathways in dermal fibroblasts and immune cells, we divided the list of rhythmic genes into four groups based on their peaking time (Method Section 5.3.8): day (ZT3 - ZT9), evening (ZT9 - ZT15), night (ZT15 - ZT21), and morning (ZT21 - ZT3 of the next day). The rhythmic genes are roughly evenly split: in fibroblasts, 426 peak during the day, 554 peak in the evening, 545 peak at night, and 421 peak in the morning; in immune cells, 129 peak during the day, 111 peak in the evening, 87 peak at night, and 105 peak in the morning. We then performed Gene Ontology analysis on the quarter-day rhythmic gene lists to identify the biological processes that are upregulated at different times of the day. We highlight some of the terms related to metabolism, signaling, cell proliferation and apoptosis, gene regulation, and immune regulation (Figure 5.5E).

During evening and night, when mice wake up, start feeding, and become active, processes such as *generation of precursor metabolites and energy*, *cellular respiration*, *mitochondrial respiratory chain complex I assembly* are upregulated in fibroblasts. Meanwhile , *glycolytic processes* are upregulated in fibroblasts, which is consistent with the finding that glycolysis is preferred at night in epidermal stem cells [107]. Additionally, similar to epidermal stem cells, more dermal fibroblasts may be in the S-phase of the cell cycle during the evening and night, as *DNA biosynthetic process* is enriched during this time. Various signaling pathways are also enriched during this time, including *prostaglandin metabolic process* and *regulation of apoptotic signaling pathway.* Gene-regulatory mechanisms such as *histone modification* and *mRNA splicing* are upregulated during the evening and night in fibroblasts. *Fibroblast migration* peaks at night, which is consistent with previous findings; mouse wounds heal fastest during the active phase [51]. Immune regulation is also circadian regulated in fibroblasts, as terms including *regulation of inflammatory response* are enriched during this time. Compared to dermal fibroblasts during the evening and night, fewer terms related to metabolism, signaling, and gene regulation are enriched in dermal immune cells. But, almost all immune regulation terms such as *defense response to virus* and *interferon-beta production* are upregulated in dermal immune cells during the evening and night, potentially contributing to shorter healing duration for wounds occurring during mice's active phase as well [51]. Additionally, such findings in mice imply that circadian regulation of immune response may be related to more severe symptoms of inflammatory skin diseases, such as psoriasis, in the evening and at night [33, 28].

In the morning and during the day, mice sleep and thus experience starvation. Consistently, rhythmic genes peaking during this time in fibroblasts enrich for *lipid catabolic process*, *glucose metabolic process*, *lipid storage*, and *response to starvation.* Interestingly, *extracellular matrix organization* and *cell-matrix adhesion* peak during the day, possibly preparing for *fibroblast migration* which peaks in the evening. For immune cells, rhythmic genes peaking during the morning and day generate fewer terms than the ones peaking during the evening and night, especially in the immune regulation category.

In summary, we found that more genes are collectively rhythmic in fibroblasts than in immune cells, while only a few rhythmic genes are shared. Additionally, more metabolism processes are diurnally regulated in fibroblasts, with respiration peaking during the evening and night, and starvation and lipid storage peaking during the morning and day. On other hand, immune regulation is almost exclusively upregulated by rhythmic genes that peak during the evening and night in immune cells.

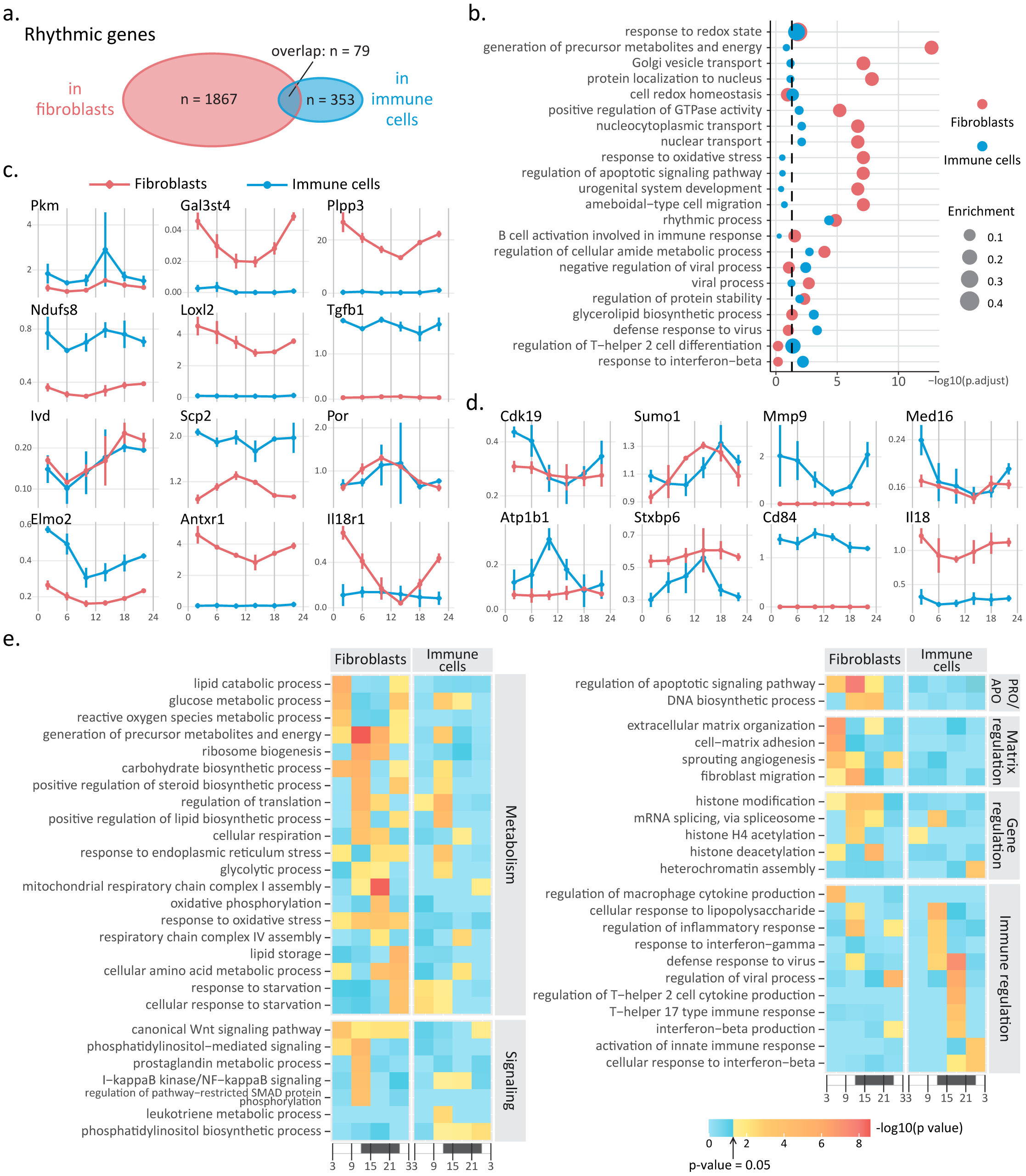


Figure 5.5: **a** JTK_Cycle identified 1946 rhythmic genes in dermal fibroblasts (red) and 432 rhythmic genes in dermal immune cells (blue). Only 79 rhythmic genes are shared by the two cell types. **b** Gene Ontology analysis performed on rhythmic genes in fibroblasts (red) and immune cells (blue) reveals divergent biological processes being diurnally regulated in the two cell types. Dot size represents enrichment score. The vertical dashed line marks adjusted p-value = 0.05. **c, d** Expression of some of the rhythmic genes found in fibroblasts (**c**), and immune cells (**d**). **e** A heatmap showing p-values for some of the biological processes enriched by rhythmic genes peaking during each quarter-day time range. Color represents p-value. Blue: insignificant; yellow to red: significant with red representing lower p-value. x-axis represents time, with white being day and black being night.

5.2.7 tauFisher determines that circadian phases are more heterogeneous in dermal immune cells than in fibroblasts.

Analysis of the pseudobulk data generated from dermal fibroblasts and immune cells reveals dampened amplitude of core clock genes (Figure 5.4D), and finds fewer rhythmic genes in immune cells than in fibroblasts (Figure 5.5A). This could mean that each individual immune cell harbors weaker circadian clock, and/or the immune cells have more heterogeneous phases, so collectively they display a dampened clock.

To look into the cause behind the dampened clock in dermal immune cells, we executed a bootstrapping approach that incorporates tauFisher for its ability to predict circadian time for transcriptomic data at different scales (Figure 5.7A). Since the heterogeneity of a set of heterogeneous clocks should be captured at any given time point, the pipeline compares circadian heterogeneity in different cell types within each time point. At each time point, we select the cell types to be examined in the scRNA-seq data and limit the expression matrix so that it only contains the predictor genes identified in the training data. To remove influence from the difference in the number of cells, we randomly sample the same number of cells for each cell type, summing the transcript counts for each gene to create a pseudobulk data set. We repeat the random sampling process to create pseudobulk replicates for each cell type. tauFisher then predicts circadian time for the pseudobulk replicates. Since the prediction outcomes are circular data, we then perform Rao's Tests for Homogeneity to compare the mean, and Wallraff Test of Angular Distances to compare the dispersion around the mean.

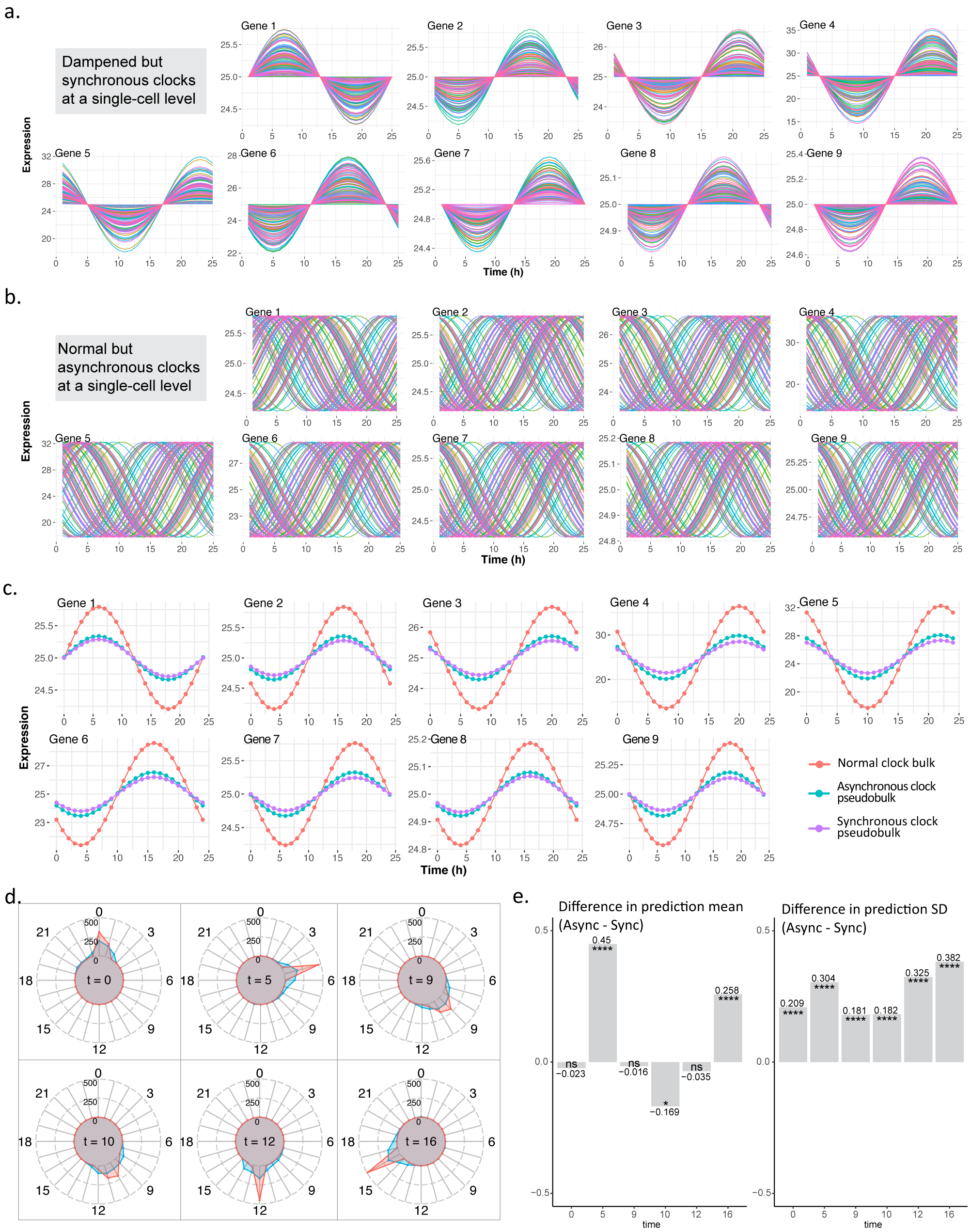


Figure 5.6: We tested tauFisher's ability to determine circadian phase heterogeneity on simulated data. **a** Simulated expression of nine predictor genes in 100 single cells that have synchronous circadian clocks but dampened amplitudes. Each line represents gene expression a cell over time. **b** Simulated expression of nine predictor genes in 100 single cells that have asynchronous circadian clocks but regular amplitudes. Each line represents gene expression a cell over time. **c** At the bulk level, scenarios in **a** and **b** generate similar patterns, which are oscillations with dampened amplitudes but similar peaking time when compared to a group of cells with normal clocks. **d** We combined tauFisher with bootstrapping and randomly selected six time points to do the comparison. We generated 500 time predictions and plotted the distributions for the cells in **a** (red) and **b**(blue). **e** Bar plots showing the differences between the prediction mean (left) and standard deviation (right) at each time point (cells in **b** - cells in **a**). ns: p-value > 0.05, $*$: p-value ≤ 0.05, p-value ≤ 0.01, $***$: p-value ≤ 0.001, $****$: p-value ≤ 0.0001.

To ensure that the pipeline works as expected, we generated simulated single-cell circadian gene expression data sets to represent a group of synchronized but dampened clocks, and a group of out-of-phase but robust clocks (Figure 5.6). As expected, the prediction outcome for the out-of-phase clocks has a significantly greater dispersion around the mean, indicating a more heterogeneous prediction outcome for the group of cells that are out-of-phase from each other (Figure 5.6).

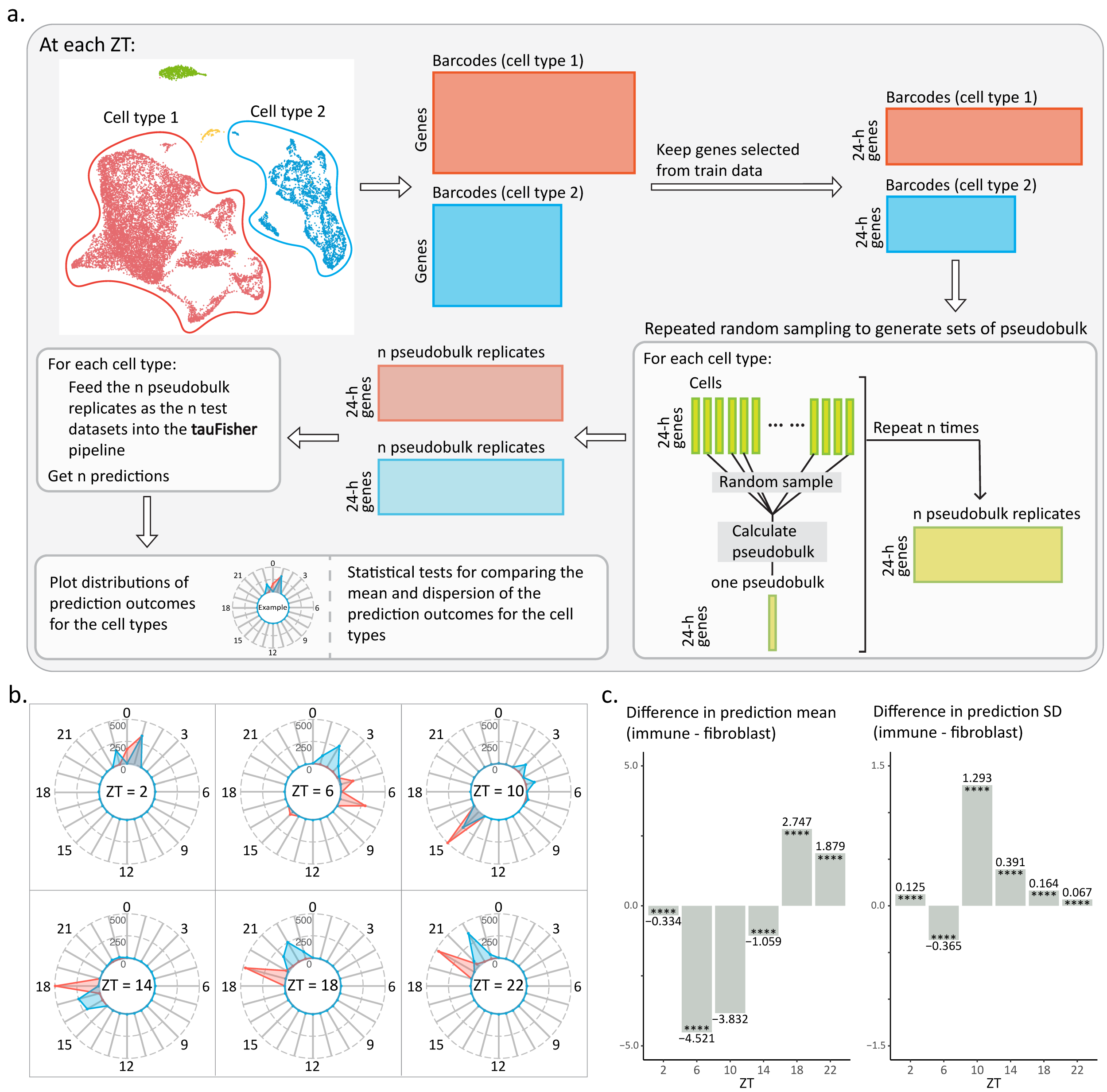


Figure 5.7: **a** A general overview of the bootstrapping approach we took to gain insight into circadian heterogeneity in phase in different cell types. **b** Radar plots showing the distribution of the prediction outcome for 500 pseudobulk replicates from dermal fibroblasts (red) and immune cells (blue). **c** Bar plots showing the differences between the prediction mean (left) and standard deviation (right) at each ZT (immune cells - fibroblasts). $***$*: p-value ≤ 0.0001

We then perform the pipeline on the scRNA-seq data we collected, focusing on the fibroblasts and immune cells. At each time point, we generated 500 pseudobulk replicates for each cell type and used tauFisher to predict the circadian time for each replicate. We then compared the distribution of the prediction outcomes from the two cell types at each time point. In general, the prediction means are centered at different times for fibroblasts and immune cells (Figure 5.7B), but around the predicted time for the whole-sample pseudobulk data (Figure 5.3C). Whether one cell type's circadian clock is ahead of the other is inconclusive (Figure 5.7C). Additionally, the distributions of the prediction outcome for immune cells are usually multimodal and not as centered as the prediction distribution for fibroblasts (Figure 5.7B). Indeed, the standard deviation in the prediction distribution is significantly greater for five out of the six ZTs, indicating that immune cells have a more heterogeneous clock phase than fibroblasts.

In summary, we were able to use tauFisher to obtain insight into the circadian heterogeneity for different cell types by predicting the circadian time for random samples from each of the cell types. We hypothesize that the circadian clock is more heterogeneous in dermal immune cells than in dermal fibroblasts, and such heterogeneity may be the reason behind the dampened core clock and fewer rhythmic genes we found in immune cells based on collective, cell-type level, gene expression data. Such a result is not unexpected, as the fibroblasts may be more homogeneous in their biological function than the immune cells, which contain dendritic cells as well as different types of macrophages and lymphocytes (Figure 5.8) that serve different immune functions. Unfortunately, we did not capture enough cells for each specific immune cell types in the scRNA-seq experiment to generate reliable pseudobulk data that is required for further circadian analysis (Figure 5.8).

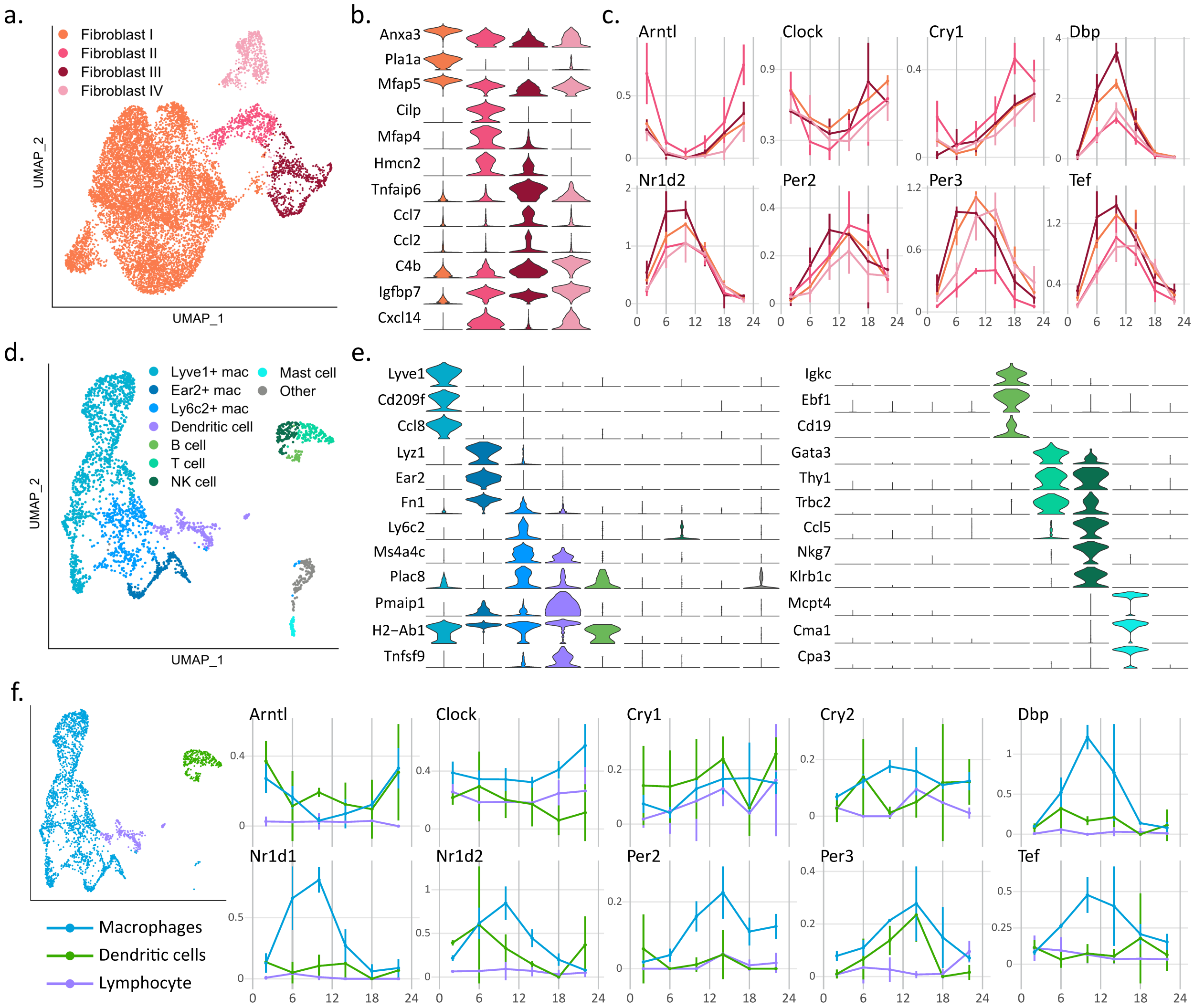


Figure 5.8: Subclustering results for the mouse dermal fibroblasts and immune cells. **a** Four subclusters were identified for the dermal fibroblasts. **b** Violin plots show expression of marker genes for each cluster. **c** Core clock gene expression for the four fibroblast subclusters show similar rhythmic pattern. **d** Eight subclusters including both myeloid cells and lymphoid cells. **e** Violin plots show expression of marker genes for each cluster. **f** Due to the low cell count for some time points, we combined the eight subclusters of immune cells form three major clusters. The core clock gene expression over time is plotted for the three major clusters. With the great variability present in the data, probably contributed by low cell counts, the core clock genes do not show robust rhythms.

5.3 Methods

5.3.1 tauFisher

tauFisher is a platform-agnostic method that predicts the circadian time from a single transcriptomic sample. The method consists of three main steps: (1) identifying a subset of diurnal genes with a period length of 24 hours, (2) calculating the difference in expression for each pair of genes, and (3) linearly transforming the differences using PCA and fitting a multinomial logistic regression on the first two principal components.

Averaging the expression matrix

The first step in tauFisher is to average each transcriptome by its genes such that the averaged gene expression matrix consists only of unique genes. The subsequent training data should consist of a gene expression matrix $X \in R^{N \times P}$ with N unique genes and P samples with known time and a vector $\tau \in R^{P}$ of the corresponding time for each sample. For scRNA-seq data, this averaged transcriptome over all the cells will be referred to as pseudobulk data.

Identifying the periodic genes

tauFisher specifies either the Lomb-Scargle [41] or JTK_Cycle [54] method in the `meta2d` function from R package MetaCycle [130] to determine the periodic genes and then selects the top ten statistically significant genes with a 24-hour period. These periodic genes are then combined with the core clock genes that also have a period of 24 hours to create the set of predictor genes M. The core clock genes for consideration in tauFisher are: *Bmal1*, *Dbp*, *Nr1d1*, *Nr1d2*, *Per1*, *Per2*, *Per3*, *Cry1*, and *Cry2*.

Subset and transform the data

Subsetting the averaged expression matrix X on the set of predictor genes M yields averaged gene expression matrix $X' \in R^{M \times P}$ with M periodic genes and P samples with known time; the vector τ should stay the same. The matrix X' is then log-transformed element-wise: $X' = \log_2(X' + 1)$.

Run Functional Data Analysis

Since experiments have different sampling intervals throughout a circadian cycle, tauFisher uses functional data analysis (FDA) to represent the discrete time points as continuous functions. This allows tauFisher to evaluate and predict the circadian time of the new samples at any time point and reduces the noise from the training samples.

Briefly, each gene m has a log-transformed measurement at discrete time points $t_1, ..., t_P \in \tau$ that are generally equally spaced but may not be. These discrete values are converted to a function Z_m with values $Z_m(t)$ for any time t using a Fourier basis expansion:

$$Z_m(t) \approx \sum_{k=1}^{K} c_{mk} \phi_k(t)$$

where $\phi_k(t)$ is the k-th basis function for $k = 1, ..., K$ and $\forall t \in \tau$, and c_{mk} is the corresponding coefficient. The Fourier basis is defined by $\phi_0(t) = 1$, $\phi_{2r-1}(t) = \sin(r\omega t)$, and $\phi_{2r}(t) = \cos(r\omega t)$ with the parameter ω determining the period $2\pi/\omega$. Since the log-transformed data matrix X' is non-negative, a positive constraint is imposed such that the positive smoothing function is defined as the exponential of an unconstrained function: $Y_m(t) = e^{Z_m(t)}$. The smoothing function also contains a roughness penalty to prevent overfitting. In practice, tauFisher sets the number of basis functions to $K = 5$ because that produces curves that are the most sinusoidal however users can specify a different number of basis functions if they wish.

Although FDA represents the discrete time points as continuous functions for each gene, tauFisher predicts circadian time at user-defined time units instead of continuous time between 0 to 24. By default, the time units are set to discrete hours. The fitted functions $Y_m(t)$ are evaluated at the user-defined time units to create the smoothed expression matrix $Y \in R^{M \times T}$, where T is the number of evaluated time points, and the new set of time points $\tau_F \in R^T$. If the time course duration of the samples spans less than 24 hours, then the fitted curves are evaluated hourly from $[0, 23]$ such that $T = 24$ to ensure all 24 hours are evaluated. If the time course duration of the samples spans greater than 24 hours, then fitted curves are evaluated from $[min(\tau), max(\tau)]$ such that $T = max(\tau) - min(\tau) + 1$.

Calculate differences between each gene pair

tauFisher then calculates the differences in the smoothed expression matrix Y for each pair of genes across time. Pairs of the same gene (i.e., the differences between Gene a and Gene a) are removed from the data matrix. Furthermore, tauFisher assumes that the effect of the difference between Gene a and Gene b and the effect of the difference between Gene b and Gene a on the predictor are not equal; thus, both pairs are included as covariates. For each time point $t_1, ..., t_T \in \tau_F$, these differences are then scaled to be $[0, 1]$.

Regress on the principal components

The differences matrix is projected onto a lower dimensional space via principal components analysis (PCA), and the first two principal components become covariates x_{i1} and x_{i2} for observation i in the multinomial regressor:

$$\log\left[\frac{P(\tau_{Fi} = t | x_{i1}, x_{i2})}{P(\tau_{Fi} = 0 | x_{i1}, x_{i2})}\right] = \beta_{t0} + \beta_{t1} x_{i1} + \beta_{t2} x_{i2}$$

All time points $t_1, ..., t_T \in \tau_F$ are converted to be $[0, 23]$, since time 0 is equal to time 24. Time zero, $\tau_F = 0$, is set as the reference level in the model. The fitted multinomial regression model is then used to predict the circadian time of the new samples.

5.3.2 Calculating the difference in predicted time and true time

To evaluate the performance of tauFisher, we need to calculate how close the predicted time is to the true time. Since the outcome is cyclic and ranges from $[0, 23]$, we apply the following conversion to calculate the true difference D from the difference d between the predicted time and true time:

$$D = \begin{cases} d - 24, & \text{if } d > 12 \\ d + 24, & \text{if } d < -12 \end{cases}$$

5.3.3 Benchmark Training

Since tauFisher relies on the `meta2d` function in the MetaCycle R package [130] to identify the periodic genes, if a training set does not include consecutive samples, we modify the input training data to have `NA`s for those samples. This may occur when the training set includes time points such as $\{4,\ 12,\ 16,\ \cdots\ \}$ and the "missing" value, 8, is assigned to the testing set. We then insert a column for the "missing" value, 8, in the input data set for tauFisher and assign `NA`s to all genes in that column. For the kidney and liver bulk sequencing data sets [137], the `meta2d` function identifies a large number (> 20) of significant genes with a period of 24 hours. To reduce the number of genes, we apply the following filters in order: (1) filter out any genes that end with "Rik" or "-ps" or begin with "MT-" or "Gm", (2) has an amplitude greater than the mean amplitude, and (3) has an amplitude greater than the median amplitude.

5.3.4 Simulating scRNA-seq circadian gene expression data sets

In Section 5.2.7, we verify that our pipeline can be used to investigate heterogeneous circadian phases using simulated scRNA-seq circadian gene expression data. We simulate two groups of data to represent two possible reasons for dampened expression: (1) a group of synchronized but dampened clock genes and (2) a group of normal (robust) but asynchronous clock genes. For both groups, the expressions of 9 representative "core clock" genes over a time course of 24-h are simulated using the following sine function:

$$y = A \sin\left(B(x + C)\right) + D$$

where A is the amplitude, C is the phase shift, D is the vertical shift, the period is $2\pi/B$, and x is a sequence of integers from 0 to 23. We set B to be 24, such that the period is $2\pi/24$, and D to be 25 to ensure that there are no negative gene expression values.

Previously, in Section 5.2.2, we used JTK_Cycle [54] to identify the periodic genes of several data sets to benchmark tauFisher; part of JTK_Cycle's output is a data frame containing inferred amplitudes and phase shifts for each gene. So, as inputs for our simulated data sets, we select the inferred amplitude and phase shift values for core clock genes *Bmal1*, *Dbp*, *Nr1d1*, *Nr1d2*, *Per1*, *Per2*, *Per3*, *Cry1*, and *Cry2* from [115]. Then, for each data set in Group (1), we simulate the expression of gene i as follows:

$$y_i = (A_i \times R_i) \sin\left(B(x + C_i)\right) + D$$

where R_i is one draw from a Beta$(1, 2)$ distribution and all other parameters are as previously stated. Similarly, for each data set in Group (2), we simulate simulate the expression of gene

i as follows:

$$y_i = A_i \sin\left(B(x + C_i + R_i)\right) + D$$

where R_i is one draw from a Normal$(0, 12)$ distribution and all other parameters are also as previously stated. We generated 100 data sets for each group, which can be thought of as the simulated expression of 9 genes for 100 single cells over 24-h.

Since tauFisher is currently used for pseudobulk expression, we then convert our simulated single cell data sets to two groups of 500 pseudobulk data sets (one for each case). We randomly select 6 time points without replacement over the course of the 24-h and use the same 6 time points for the simulated pseudobulk data sets. For each "gene", we randomly select 20% of the "single cells" without replacement and sum their expression to obtain a vector of pseudobulk replicates for that gene over the 6 selected time points. We repeat this procedure 500 times for each group, generating 500 pseudobulk data sets per group.

5.3.5 scRNA-seq experiments

Mouse strains, husbandry

Wild type male C57BL/6 were housed under 12:12 light-dark cycle for two weeks prior to and during the time of experiment. To collect telogen skin, mice were about 54 days old by the time of sample collection.

Sample collection and sequencing

Immediately after sacrificing a mouse with CO_2, hair on dorsal skin was removed with an electric razor and Nair Hair Removal cream. After the dorsal skin is isolated from the body, fat and remaining blood vessels were scrapped away. A circular piece of skin was obtained with a 12mm biopsy punch, and minced into tiny pieces. 1x collagenase was then added to the minced skin and the suspension was incubated at 37 °C for 1.5 hours. The suspension is then filtered with a $70\mu m$ and a $40\mu m$ cell strainers to obtain single cells. SYTOX blue viability dye was then added to the cell suspension and live cells were sorted out using FACS at the UCI Institute for Immunology Flow Cytometry Facility.

Samples were collected every four hours for three days to generate in total 18 sample. The Chromium Single Cell 3' v3 (10x Genomics) libraries were prepared and sequenced by University of California Irvine Genomic High Throughput Facility with Illumina NovaSeq6000.

5.3.6 Preprocessing for benchmarking

For GSE56931, we subset the data set provided in the TimeSignatR [10] package to only include the 24-hour normal baseline time points and filtered out the 38 hours of continuous wakefulness and subsequent recovery sleep.

For GSE38622, expression matrix was normalized as explained in [39].

For GSE157077, we used the transcriptomes of the mice who were fed normal chow through an entire circadian cycle (24 hours). Since the mice were maintained on a 12-hour light/12-hour dark cycle, we chose to concatenate the three replicates of 24 hours each to create one set of samples over 72 hours.

For GSE54650, the raw CEL files for the kidney and liver were imported using the function `read.celfiles` in R package oligo. Each raw data matrix was then normalized with Robust Multiarray Average (RMA) using the function `rma`. To map the GPL6246 platform ID_REF to Ensembl transcript IDs, we used the transcript cluster ID and gene assignments listed in the table provided at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL6246`

For each transcript cluster ID, we removed all gene assignments unless they were Ensembl transcript IDs or started with Gm. If a transcript cluster ID was mapped to more than one gene, then we replicated that row by the number of genes (e.g., transcript reference ID 10344614 is assigned to three Ensembl transcript IDs so that row in the normalized data set was replicated three times). The expression for each transcript cluster ID was then divided by the number of genes assigned (e.g., since transcript reference ID 10344614 has three gene assignments, the values for all three rows in the normalized data set were divided by three). Transcript cluster IDs that were not assigned to any genes were removed from the normalized data set. To convert the Ensembl transcript IDs to gene names, we use R package biomaRt [29, 30]. If biomaRt did not find a gene name, then we kept the original Ensembl transcript ID.

For GSE54651, we converted the Ensembl gene IDs to gene names using R package biomaRt [29, 30]. If biomaRt did not find a gene name, then we kept the original Ensembl gene ID.

For each time point in GSE117295 and the scRNA-seq data we collected in this study, we summed the counts of each gene in all the cells without any pre-processing to create a pseudobulk data set. In the case where the same gene occurs multiple times in the data, we took the mean of those entries. The resulting pseudobulk data at each time point is a single row vector in which each entry represents the expression value of a unique gene. The light-stimulated group is not considered in this paper.

5.3.7 scRNA-seq data analysis for dermal skin

We used Cell Ranger version 3.1.0 with MM10 reference to process the raw sequencing output. The downstream analysis was done in Seurat V3 according to the vignette.

Cell with 850-7800 features and less than 13% of mitochondrial genes were kept. The SC-

Transform function was performed on each sample and 3250 integration features were selected using SelectIntegrationFeatures for each sample. Principal component analysis was then done on the integrated data set and the Louvian algorithm was used to generate the clusters. Cluster identities were then determined in combination of marker genes found in the current clustering outcome and feature plots of canonical marker genes.

5.3.8 Pseudobulk data analysis for dermal skin

meta2d from the MetaCycle package was used on the pseudobulk data generated from the scRNA-seq data collected from dermal skin to identify rhythmic genes. Genes with JTK_pvalue<0.05 were determined to be significantly rhythmic. We used the meta2d_phase column to split the rhythmic genes into four groups based on their peaking time.

Gene Ontology analysis was performed using ClusterProfiler in R with p-value < 0.05 as the significance cutoff.

5.3.9 Statistics for circular data

The "circular" package was used to perform statistical calculations and tests, including calculation of the mean and standard deviation, as well as the Rao's Tests for Homogeneity and the Wallraff Test of Angular Distances, for the circadian time prediction output in Section 5.2.7.

5.4 Discussion

Circadian time determination is a key step in the implementation of circadian medicine. To maximize effectiveness while minimizing side effects of treatments, it is necessary to take into consideration the patient's and relevant tissue's actual circadian time. For example, on-pump cardiac surgeries in the afternoon are less likely to cause perioperative myocardial injury than when conducted in the morning [81] and cancer radiation therapy in the morning causes less skin damage than in the afternoon [105]. There have been several predictors of circadian time for patients and organs based on transcriptomic data [10, 55, 1], but to ensure wide applicability of this approach, a sequencing platform-agnostic method requiring low number of testing samples is desired. In this study, we developed tauFisher, a computational pipeline that accurately predicts circadian time from a single transcriptomic data set applicable to within-platform and cross-platform training-testing scenarios. Most importantly, tauFisher trained on bulk RNA-seq data sets accurately predicts time labels for scRNA-seq data, enabling the study and comparison of the circadian clocks at the cell-type level.

Once trained, tauFisher requires a single sample from the test subject to predict the circadian time. We examined tauFisher's ability to predict circadian time when the training and test data are from the same study, and we benchmarked it against state-of-the-art methods ZeitZeiger [55] and TimeSignatR [10]. tauFisher performed the best in accuracy and RMSE for almost all data sets. ZeitZeiger failed to run for several of the data sets due to linear dependency issues. When it did run, ZeitZeiger achieves similar accuracy and RMSE as tauFisher, possibly because both predictors build on principal component analysis, suggesting that the molecular clock is well captured and represented by orthogonal linear combinations of feature genes. TimeSignatR performed the worst in the benchmark, even when a second sample was inferred. For time inference from a single sample, we reported worse general performance for TimeSignatR than reported in [10]. This could be due to the difference in the training data. In [10], TimeSignatR was trained on an independent data set (GSE39445) from a different study [80] before testing on GSE56931 and GSE113883. Here we trained and tested TimeSingatR within GSE56931 and GSE113883. [10] also reported more accurate predictions by TimeSignatR with 2-point calibration than with just 1-point. In this study's benchmarking, TimeSignatR with 2-point calibration performed worse. Such discrepancy in performance reported between this study and [10] could be due to our implementation of the two-point calibration step in the TimeSignatR. In their vignette, the authors use the entire data set (train and test data) during this two-point calibration step instead of just the training set. As tauFisher and ZeitZeiger do not utilize the test data in the processing of the train data, to make a fair comparison of their performance in predicting circadian time from one single test sample, we do not make the test sample available to the predictor until the actual testing step when implementing all three methods.

One of the most powerful features of tauFisher is its ability to accurately predict circadian time when trained and tested on data sets collected from different assay platforms under different experimental settings. tauFisher achieves high accuracy and low RMSE in not only bulk-to-bulk cross-platform predictions, but also bulk-to-scRNA-seq predictions. The consistency in performance despite drastically different assay methods and experimental setups suggests that tauFisher captures and extracts the underlying biological correlations in gene expressions while minimizing the effects of the noise and variability introduced by subjects and technology.

Two key steps in tauFisher help achieve this: functional data analysis for training data and within-sample normalization for both training and test data. Functional data analysis for the training data enables tauFisher to remove minor noise, smooth the time expression curves, and generate the expression data between the sampled time points. The within-sample normalization step for both training and test data calculates the difference between each pair of predictor genes at a given time point so that the feature matrix is expanded while some baseline noise is removed. The differences between the genes are then re-scaled to be between 0 and 1 so that the data become unit-less. Doing so in parallel on the training and test data sets brings them individually to the same scale, instead of batch-correcting the train to the scale of the test or the opposite. This allows testing of independent data sets without re-training. We note that this within-sample normalization is different from the within-subject normalization in [10], which is based on mean expression calculated from multiple samples collected from more than one time point over the circadian cycle.

In addition to testing tauFisher on published data sets, we also collected a time series of scRNA-seq from mouse dermis. Consistent with previous findings [88, 28], the circadian rhythm is robustly present in the dermis and the oscillatory patterns of the core clock genes agree with published data [39]. Comparing the rhythmic genes in fibroblasts and immune cells, we found that many pathways and processes are rhythmically regulated in a cell type-specific manner. While we only collected dermal cells, the circadian clock is present in all skin layers, and the circadian clock regulates hair cycles for both mouse and human [28, 88, 69, 2]. A time series of scRNA-seq data containing all skin cells will be particularly useful for investigating and comparing the circadian clocks in different skin layers and hair follicles, as well as exploring whether cell-cell communication is clock-controlled in the skin.

Combining tauFisher with other methods can guide the application of circadian medicine by providing additional insights and explanation of clinically observed circadian dysfunction. Dampened clock gene expressions have been observed in psoriasis-affected skin [42, 135], as well as in various types of cancer [132, 109, 22, 62, 52, 56], including melanoma, head and neck squamous cell carcinoma, breast cancer, and colorectal cancer. There is also evidence that restoring dampened circadian oscillations in diseased tissues can be effective. Dexamethasone, a glucocorticoid that activates and synchronizes the clock between cells, restores the rhythmicity of cell cycle and reduces proliferation and growth of melanoma and colon carcinoma cells [62]. Seliciclib, a cyclin-dependent kinase inhibitor that indirectly interact with the circadian clock, restores clock gene expression and cell cycle rhythms in

Glasgow osteosarcoma tumors and reduces tumor growth [56].

There are two possible behind-the-scene causes of dampened circadian rhythms at a bulk level: First, the circadian rhythm is dampened in every cell, but the cells are synchronous to each other. Second, the clock is normally functioning in every cell, but the cells are out of phase relative to each other. Understanding which of the two scenarios is responsible for a dampened bulk-level clock gene expression is particularly important because in one case, it would be optimal to stimulate the clock to restore the circadian clock in the diseased tissue, while in the other case, synchronizing the clock is more suitable.

Here, we observed that the collective circadian rhythm in dermal immune cells is dampened compared to fibroblasts. We incorporate tauFisher with bootstrapping to investigate the cause behind the dampened collective circadian rhythm in dermal immune cells. tauFisher's prediction outcome suggests that the circadian phases in dermal immune cells are more heterogeneous than those in dermal fibroblasts, and this heterogeneity may contribute to the dampened rhythm in immune cells at a collective level. While we will need to perform live cell imaging with reporters for core clock genes to validate this hypothesis, tauFisher's success in simulated data is reassuring.

Besides being useful in circadian medicine, tauFisher opens up new possibilities in circadian research. For example, it expands the scRNA-seq databases for circadian research by adding (circadian) time labels to existing scRNA-seq data sets. Beyond scRNA-seq data, adding time labels for all existing transcriptomic data sets is important since the clock modulates many protein coding genes and it is necessary to know whether a significant gene is truly differentially regulated by a condition or the expression appears to be different because the samples were collected at different times. Additionally, combining tauFisher with a batch-effect correction method may facilitate a cleaner integration and help minimize the effect of the circadian clock in transcriptomic data analysis. This approach harbors great potential as many efforts are going into integrating data sets from different studies to create meta databases such as in the Human Cell Atlas.

In summary, tauFisher's consistent and robust performance in accurately predicting circadian time from a single transcriptomic data makes it a useful addition to the toolbox of circadian medicine and research.

5.5 Data availability

All published data sets used in this paper can be accessed through their respective GEO accession codes. Although the data sets in [6] and [10] are both accessible through their GEO accession codes, this paper used the versions provided in the TimeSignatR package [10]: `https://github.com/braunr/TimeSignatR`.

The time series of scRNA-seq data from mouse dermal skin will be made available to the GEO database.

5.6 Code availability

tauFisher is available as an R package, which is available at `https://github.com/micnngo/tauFisher` (note that this will be a private repository until publication).

The two methods we compared tauFisher against are also available as R packages: TimeSignatR at `https://github.com/braunr/TimeSignatR` and ZeitZeiger at `https://github.com/hugheylab/zeitzeiger`. For TimeSignatR, we modified the vignette the authors provided so that the two point calibration only calibrates the training data set; the original vignette calibrates the concatenated training and test data sets.

5.7 Acknowledgements

This project is supported by an award entitled "Skin Biology Resource-Based Center at UCI - Systems Biology Core" from the Center of Complex Biological Systems funded by NIH/NIAMS P30-AR075047 (J.D. and M.N.); National Institute of Health grants R01AR056439 and P30AR075047 (B.A.); NSF grant DMS1763272 and a grant from the Simons Foundation (594598) (J.D., M.N. and B.A.); National Institute of General Medical Sciences, National Research Service Award GM136624 (J.D.); NSF grant DMS1936833 (M.N. and B.S.); and the California Institute for Regenerative Medicine Training Program Award EDUC4-12822 (S.S.K.).

Chapter 6

Conclusion

As sequencing technologies advances, leveraging latent representations will become more crucial to analyze the data from these platforms and investigate biological processes of interest. Currently, many platforms are already able to sequence millions of cells and trying to do a traditional biclustering or regression on that data matrix would require a large amount of time and computational resources. Furthermore, at the single cell resolution, the data is extremely noisy; teasing out the true signal and reducing the amount of noise are two other crucial aspects in analyzing single cell data. Despite these computational challenges, the popularity of single cell sequencing has contributed to many advances in biology, such as the ability of interrogating cell lineages, examining cell heterogeneity and heterogeneous responses to a stimulus.

To this end, we have presented two computational frameworks leveraging the latent representations of single cell data to investigate clonal hematopoiesis and circadian rhythms. In the former, we utilize a Bayesian approach to identify significant differentially expressed genes between two samples and rank these groups of genes. We then use these gene groupings to inform our cell clusterings to identify heterogeneous subpopulations within the hematopoietic stem and progenitor cells. With this approach, we are able to confirm that a MPN patient is more likely to have a higher inflammatory response to stimulation than an unaffected patient.

In the latter, we present a pipeline that also uses the differences between two groups (in this case, between each pair of rhythmic genes) to predict the circadian time of the sample of interest. This pipeline is beneficial to circadian medicine, particularly precision medicine, and circadian research. The ability to estimate a tissue sample's circadian time can help doctors cater their patients' treatments to minimize discomfort or maximize efficacy, and this also expands current databases for circadian research by incorporating time labels.

Both methods are limited by technology among others however. For the clustering, computational efficiency is still a concern as well as the robustness. We have not performed a thorough comparison between each step and hyperparameter. For the circadian pipeline, we are heavily reliant on the selection of rhythmic genes. Further benchmarking and refactoring will be necessary.

Bibliography

[1] F. Agostinelli, N. Ceglia, B. Shahbaba, P. Sassone-Corsi, and P. Baldi. What time is it? deep learning approaches for circadian rhythms. *Bioinformatics*, 32(12):i8–i17, jun 2016.

[2] Y. Al-Nuaimi, J. A. Hardman, T. Bíró, I. S. Haslam, M. P. Philpott, B. I. Tóth, N. Farjo, B. Farjo, G. Baier, R. E. Watson, B. Grimaldi, J. E. Kloepper, and R. Paus. A meeting of two chronobiological systems: Circadian proteins period1 and BMAL1 modulate the human hair cycle clock. *Journal of Investigative Dermatology*, 134(3):610–619, mar 2014.

[3] S. Allain-Maillet, A. Bosseboeuf, N. Mennesson, M. Bostoën, L. Dufeu, E. H. Choi, C. Cleyrat, O. Mansier, E. Lippert, Y. L. Bris, J. M. Gombert, F. Girodon, M. Pettazzoni, E. Bigot-Corbel, and S. Hermouet. Anti-Glucosylsphingosine Autoimmunity, JAK2V617F-Dependent Interleukin-1β and JAK2V617F-Independent Cytokines in Myeloproliferative Neoplasms. *Cancers*, 12(9):1–24, 9 2020.

[4] T. S. Andrews and M. Hemberg. Identifying cell populations with scRNASeq. *Molecular Aspects of Medicine*, 59:114–122, 2018.

[5] T. S. Andrews, V. Y. Kiselev, D. McCarthy, and M. Hemberg. Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data, 2021.

[6] E. S. Arnardottir, E. V. Nikonova, K. R. Shockley, A. A. Podtelezhnikov, R. C. Anafi, K. Q. Tanis, G. Maislin, D. J. Stone, J. J. Renger, C. J. Winrow, and A. I. Pack. Blood-Gene Expression Reveals Reduced Circadian Rhythmicity in Individuals Resistant to Sleep Deprivation. *Sleep*, 37(10):1589–1600, 10 2014.

[7] J. Bass. Circadian topology of metabolism. *Nature*, 491:348–356, 11 2012.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[9] R. L. Bowman, L. Busque, and R. L. Levine. Clonal Hematopoiesis and Evolution to Hematopoietic Malignancies. *Cell stem cell*, 22(2):157, 2 2018.

[10] R. Braun, W. L. Kath, M. Iwanaszko, E. Kula-Eversole, S. M. Abbott, K. J. Reid, P. C. Zee, and R. Allada. Universal method for robust detection of circadian state from gene

expression. *Proceedings of the National Academy of Sciences*, 115(39):E9247–E9256, 9 2018.

[11] P. Brennecke, S. Anders, J. K. Kim, A. A. Kołodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni, and M. G. Heisler. Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 2013.

[12] E. D. Buhr, S. Vemaraju, N. Diaz, R. A. Lang, and R. N. V. Gelder. Neuropsin (OPN5) Mediates Local Light-Dependent Induction of Circadian Clock Genes and Circadian Photoentrainment in Exposed Murine Skin. *Current Biology*, 29(20):3478–3487.e4, 10 2019.

[13] E. D. Buhr, S. Vemaraju, N. Diaz, R. A. Lang, and R. N. V. Gelder. Neuropsin (OPN5) mediates local light-dependent induction of circadian clock genes and circadian photoentrainment in exposed murine skin. *Current Biology*, 29(20):3478–3487.e4, oct 2019.

[14] E. D. Buhr, S.-H. Yoo, and J. S. Takahashi. Temperature as a universal resetting cue for mammalian circadian oscillators. *Science*, 330:379–385, 10 2010.

[15] J. Chang and J. W. Fisher III. Parallel Sampling of DP Mixture Models using Sub-Clusters Splits. In *Proceedings of the Neural Information Process Systems (NIPS)*, 2013.

[16] X. Chen, S. A. Teichmann, and K. B. Meyer. From Tissues to Cell Types and Back: Single-Cell Gene Expression Analysis of Tissue Architecture. *Annual Review of Biomedical Data Science*, 1:29–51, 2018.

[17] Y. Cheng and G. M. Church St. Biclustering of Expression Data. 8:93–103, 2000.

[18] V. P. Čokić, O. Mitrović-Ajtić, B. B. Beleslin-Čokić, D. Marković, M. Buač, M. Diklić, N. Kraguljac-Kurtović, S. Damjanović, P. Milenković, M. Gotić, and P. K. Raj. Proinflammatory Cytokine IL-6 and JAK-STAT Signaling Pathway in Myeloproliferative Neoplasms. *Mediators of Inflammation*, 2015, 2015.

[19] L. M. Coussens and Z. Werb. Inflammation and cancer. *Nature*, 420(6917):860, 12 2002.

[20] K. H. Cox and J. S. Takahashi. Circadian clock genes and the transcriptional architecture of the clock mechanism. *Journal of Molecular Endocrinology*, 63:R93–R102, 11 2019.

[21] B. Craver, Q. Nguyen, G. Ramanathan, and A. G. Fleischman. Single-Cell RNA-Seq to Assess Differential Responses to Tnfα in Human Hematopoietic Stem and Progenitor Cells in Myeloproliferative Neoplasm. *Blood*, 134(Supplement_1):2518, 11 2019.

[22] K. Davis, L. C. Roden, V. D. Leaner, and P. J. van der Watt. The tumour suppressing role of the circadian clock. *IUBMB Life*, 71(7):771–780, jan 2019.

[23] F. Denti, R. Azevedo, C. Lo, D. Wheeler, S. P. Gandhi, M. Guindani, and B. Shahbaba. A Horseshoe mixture model for Bayesian screening with an application to light sheet fluorescence microscopy in brain imaging. 6 2021.

[24] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274, 2001.

[25] I. S. Dhillon, S. Mallela, and D. S. Modha. Information-Theoretic Co-clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003.

[26] B. A. Dickerman, S. C. Markt, M. Koskenvuo, C. Hublin, E. Pukkala, L. A. Mucci, and J. Kaprio. Sleep disruption, chronotype, shift work, and prostate cancer risk and mortality: a 30-year prospective cohort study of finnish twins. *Cancer Causes &amp Control*, 27(11):1361–1370, oct 2016.

[27] O. Dinari, A. Yu, O. Freifeld, and J. W. Fisher III. Distributed MCMC Inference in Dirichlet Process Mixture Models Using Julia. In *19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, 2019.

[28] J. Duan, E. N. Greenberg, S. S. Karri, and B. Andersen. The circadian clock and diseases of the skin. *FEBS Letters*, 595(19):2413–2436, sep 2021.

[29] S. Durinck, Y. Moreau, A. Kasprzyk, S. Davis, B. De Moor, A. Brazma, and W. Huber. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, 8 2005.

[30] S. Durinck, P. T. Spellman, E. Birney, and W. Huber. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols 2009 4:8*, 4(8):1184–1191, 7 2009.

[31] B. Efron. Microarrays, Empirical Bayes and the Two-Groups Model. *Statistical Science*, 23(1):1–22, 2 2008.

[32] M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science (New York, N.Y.)*, 297(5584):1183–6, 8 2002.

[33] F. Ferguson, G. Lada, H. Hunter, C. Bundy, A. Henry, C. Griffiths, and C. Kleyn. Diurnal and seasonal variation in psoriasis symptoms. *Journal of the European Academy of Dermatology and Venereology*, 35(1), aug 2020.

[34] T. S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2), 1973.

[35] D. A. C. Fisher, J. S. Fowles, A. Zhou, and S. T. Oh. Inflammatory Pathophysiology as a Contributor to Myeloproliferative Neoplasms. *Frontiers in Immunology*, 12:2034, 6 2021.

[36] A. G. Fleischman, K. J. Aichberger, S. B. Luty, T. G. Bumm, C. L. Petersen, S. Doratotaj, K. B. Vasudevan, D. H. LaTocha, F. Yang, R. D. Press, M. M. Loriaux, H. L. Pahl, R. T. Silver, A. Agarwal, T. O'Hare, B. J. Druker, G. C. Bagby, and M. W. Deininger. TNFα facilitates clonal expansion of JAK2V617F positive cells in myeloproliferative neoplasms. *Blood*, 118(24):6392, 12 2011.

[37] S. Fowlkes, C. Murray, A. Fulford, T. D. Gelder, and N. Siddiq. Myeloproliferative neoplasms (MPNs) – Part 1: An overview of the diagnosis and treatment of the "classical" MPNs. *Canadian Oncology Nursing Journal*, 28(4):262, 10 2018.

[38] A. Gelman, J. B. B. Carlin, H. S. S. Stern, and D. B. B. Rubin. *Bayesian Data Analysis, Third Edition (Texts in Statistical Science)*. 2014.

[39] M. Geyfman, V. Kumar, Q. Liu, R. Ruiz, W. Gordon, F. Espitia, E. Cam, S. E. Millar, P. Smyth, A. Ihler, J. S. Takahashi, and B. Andersen. Brain and muscle arnt-like protein-1 (bmal1) controls circadian cell proliferation and susceptibility to uvb-induced dna damage in the epidermis. *Proceedings of the National Academy of Sciences of the United States of America*, 109:11758, 7 2012.

[40] M. Geyfman, V. Kumar, Q. Liu, R. Ruiz, W. Gordon, F. Espitia, E. Cam, S. E. Millar, P. Smyth, A. Ihler, J. S. Takahashi, and B. Andersen. Brain and muscle Arnt-like protein-1 (BMAL1) controls circadian cell proliferation and susceptibility to UVB-induced DNA damage in the epidermis. *Proceedings of the National Academy of Sciences of the United States of America*, 109(29):11758, 7 2012.

[41] E. F. Glynn, J. Chen, and A. R. Mushegian. Detecting periodic patterns in unevenly spaced gene expression time series using Lomb–Scargle periodograms. *Bioinformatics*, 22(3):310–316, 2 2006.

[42] E. N. Greenberg, M. E. Marshall, S. Jin, S. Venkatesh, M. Dragan, L. C. Tsoi, J. E. Gudjonsson, Q. Nie, J. S. Takahashi, and B. Andersen. Circadian control of interferon-sensitive gene expression in murine skin. *Proceedings of the National Academy of Sciences of the United States of America*, 117:5761–5771, 3 2020.

[43] G. Greenfield, M. F. McMullin, and K. Mills. Molecular pathogenesis of the myeloproliferative neoplasms. *Journal of Hematology & Oncology 2021 14:1*, 14(1):1–18, 6 2021.

[44] D. Gutierrez and J. Arbesman. Circadian dysrhythmias, physiological aberrations, and the link to skin cancer. *International Journal of Molecular Sciences*, 17(5):621, apr 2016.

[45] J. Hansen. Night shift work and risk of breast cancer. *Current Environmental Health Reports*, 4(3):325–339, aug 2017.

[46] Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M.

McElrath, C. A. Blish, R. Gottardo, P. Smibert, and R. Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 6 2021.

[47] A. Haque, J. Engel, S. A. Teichmann, and T. Lönnberg. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*, 9(75), 2017.

[48] J. A. Hartigan. Direct Clustering of a Data Matrix. *Journal of the American Statistical Association*, 67(337):123, 3 1972.

[49] T. Hashimshony, N. Senderovich, G. Avital, A. Klochendler, Y. de Leeuw, L. Anavy, D. Gennert, S. Li, K. J. Livak, O. Rozenblatt-Rosen, Y. Dor, A. Regev, and I. Yanai. CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology*, 17(77), 2016.

[50] H. C. Hasselbalch and M. E. Bjørn. MPNs as Inflammatory Diseases: The Evidence, Consequences, and Perspectives, 2015.

[51] N. P. Hoyle, E. Seinkmane, M. Putker, K. A. Feeney, T. P. Krogager, J. E. Chesham, L. K. Bray, J. M. Thomas, K. Dunn, J. Blaikley, and J. S. O'Neill. Circadian actin dynamics drive rhythmic fibroblast mobilization during wound healing. *Science Translational Medicine*, 9(415), nov 2017.

[52] C.-M. Hsu, S.-F. Lin, C.-T. Lu, P.-M. Lin, and M.-Y. Yang. Altered expression of circadian clock genes in head and neck squamous cell carcinoma. *Tumor Biology*, 33(1):149–155, nov 2011.

[53] X. Hu, J. li, M. Fu, X. Zhao, and W. Wang. The JAK/STAT signaling pathway: from bench to clinic. *Signal Transduction and Targeted Therapy 2021 6:1*, 6(1):1–33, 11 2021.

[54] M. E. Hughes, J. B. Hogenesch, and K. Kornacker. JTK_CYCLE: an efficient non-parametric algorithm for detecting rhythmic components in genome-scale datasets. *Journal of biological rhythms*, 25(5):372, 10 2010.

[55] J. J. Hughey, T. Hastie, and A. J. Butte. ZeitZeiger: Supervised learning for high-dimensional data from an oscillatory system. *Nucleic Acids Research*, 44(8):e80, 5 2016.

[56] I. Iurisci, E. Filipski, J. Reinhardt, S. Bach, A. Gianella-Borradori, S. Iacobelli, L. Meijer, and F. Le
vi. Improved tumor control through circadian clock induction by seliciclib, a cyclin-dependent kinase inhibitor. *Cancer Research*, 66(22):10720–10728, nov 2006.

[57] M. Jagannathan-Bogdan and L. I. Zon. Hematopoiesis. *Development (Cambridge, England)*, 140(12):2463, 6 2013.

[58] S. Kaiser, R. Santamaria, T. Khamiakova, M. Sill, R. Theron, L. Quintales, F. Leisch, and E. De Troyer. biclust: BiCluster Algorithms, 2018.

[59] R. Kawahara. Hematopoiesis. *xPharm: The Comprehensive Pharmacology Reference*, pages 1–5, 1 2007.

[60] D. J. Kennaway. A critical review of melatonin assays: Past and present. *Journal of Pineal Research*, page e12572, apr 2019.

[61] M. Khosravipour, M. Shahmohammadi, and H. V. Athar. The effects of rotating and extended night shift work on the prevalence of metabolic syndrome and its components. *Diabetes &amp Metabolic Syndrome: Clinical Research &amp Reviews*, 13(6):3085–3089, nov 2019.

[62] S. Kiessling, L. Beaulieu-Laroche, I. D. Blum, D. Landgraf, D. K. Welsh, K.-F. Storch, N. Labrecque, and N. Cermakian. Enhancing circadian clock function in cancer cells inhibits tumor growth. *BMC Biology*, 15(1), feb 2017.

[63] C. Kim, H. Lee, J. Jeong, K. Jung, and B. Han. MarcoPolo: a method to discover differentially expressed genes in single-cell RNA-seq data without depending on prior clustering. *Nucleic Acids Research*, 50(12):e71–e71, 7 2022.

[64] V. Kiselev and M. Hemberg. Collection of public scRNA-Seq datasets used by our group, 2017.

[65] M. Kleppe, M. Kwak, P. Koppikar, M. Riester, M. Keller, L. Bastian, T. Hricik, N. Bhagwat, A. S. McKenney, E. Papalexi, O. Abdel-Wahab, R. Rampal, S. Marubayashi, J. J. Chen, V. Romanet, J. S. Fridman, J. Bromberg, J. Teruya-Feldstein, M. Murakami, T. Radimerski, F. Michor, R. Fan, and R. L. Levine. JAK-STAT pathway activation in malignant and nonmalignant cells contributes to MPN pathogenesis and therapeutic response. *Cancer discovery*, 5(3):316–331, 2015.

[66] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein. Spectral biclustering of microarray data: Coclustering genes and conditions. *Genome Research*, 2003.

[67] V. Lang, S. Ferencik, B. Ananthasubramaniam, A. Kramer, and B. Maier. Susceptibility rhythm to bacterial endotoxin in myeloid clock-knockout mice. *eLife*, 10:e62469, oct 2021.

[68] L. Lazzeroni and A. Owen. Plaid Models for Gene Expression Data. *Statistica Sinica*, 12(1):61–86, 2002.

[69] K. K. Lin, V. Kumar, M. Geyfman, D. Chudova, A. T. Ihler, P. Smyth, R. Paus, J. S. Takahashi, and B. Andersen. Circadian clock genes contribute to the regulation of hair follicle cycling. *PLoS Genetics*, 5(7):e1000573, jul 2009.

[70] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, 15(12):550, 12 2014.

[71] M. D. Luecken and F. J. Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, 6 2019.

[72] A. Lun and D. Risso. SingleCellExperiment: S4 Classes for Single Cell Data, 2019.

[73] E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, and S. A. McCarroll. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, 2015.

[74] S. Marques, A. Zeisel, S. Codeluppi, D. van Bruggen, A. Mendanha Falcão, L. Xiao, H. Li, M. Häring, H. Hochgerner, R. A. Romanov, D. Gyllborg, A. Muñoz Manchado, G. La Manno, P. Lönnerberg, E. M. Floriddia, F. Rezayee, P. Ernfors, E. Arenas, J. Hjerling-Leffler, T. Harkany, W. D. Richardson, S. Linnarsson, and G. Castelo-Branco. Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science (New York, N.Y.)*, 352(6291):1326–1329, 6 2016.

[75] S. Masri, K. Kinouchi, and P. Sassone-Corsi. Circadian clocks, epigenetics, and cancer. *Current Opinion in Oncology*, 27(1):50–56, jan 2015.

[76] E. Masselli, G. Pozzi, G. Gobbi, S. Merighi, S. Gessi, M. Vitale, and C. Carubbi. Cytokine Profiling in Myeloproliferative Neoplasms: Overview on Phenotype Correlation, Outcome Prediction, and Role of Genetic Variants. *Cells*, 9(9), 9 2020.

[77] L. F. Mendez Luque, A. L. Blackmon, G. Ramanathan, and A. G. Fleischman. KEY ROLE OF INFLAMMATION IN MYELOPROLIFERATIVE NEOPLASMS: INSTIGATOR OF DISEASE INITIATION, PROGRESSION. AND SYMPTOMS. *Current hematologic malignancy reports*, 14(3):145, 6 2019.

[78] H. Mi, A. Muruganujan, D. Ebert, X. Huang, and P. D. Thomas. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*, 47:419–426, 2018.

[79] H. Mi, A. Muruganujan, X. Huang, D. Ebert, C. Mills, X. Guo, and P. D. Thomas. Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nature Protocols*, 14(3):703–721, 3 2019.

[80] C. S. Möller-Levet, S. N. Archer, G. Bucca, E. E. Laing, A. Slak, R. Kabiljo, J. C. Y. Lo, N. Santhi, M. v. Schantz, C. P. Smith, and D.-J. Dijk. Effects of insufficient sleep on circadian rhythmicity and expression amplitude of the human blood transcriptome. *Proceedings of the National Academy of Sciences*, 110(12):E1132–E1141, 3 2013.

[81] D. Montaigne, X. Marechal, T. Modine, A. Coisne, S. Mouton, G. Fayad, S. Ninni, C. Klein, S. Ortmans, C. Seunes, C. Potelle, A. Berthier, C. Gheeraert, C. Piveteau, R. Deprez, J. Eeckhoute, H. Duez, D. Lacroix, B. Deprez, B. Jegou, M. Koussa, J.-L. Edme, P. Lefebvre, and B. Staels. Daytime variation of perioperative myocardial injury in cardiac surgery and its prevention by rev-erb antagonism: a single-centre propensity-matched cohort study and a randomised study. *The Lancet*, 391(10115):59–69, jan 2018.

[82] C. J. Morris, T. E. Purvis, K. Hu, and F. A. J. L. Scheer. Circadian misalignment increases cardiovascular disease risk factors in humans. *Proceedings of the National Academy of Sciences*, 113(10), feb 2016.

[83] T. Mou, W. Deng, F. Gu, Y. Pawitan, and T. N. Vu. Reproducibility of Methods to Detect Differentially Expressed Genes from Single-Cell RNA Sequencing. *Frontiers in Genetics*, 10, 1 2020.

[84] R. M. Neal. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, 2000.

[85] V. A. Padilha and R. J. G. B. Campello. A systematic comparative evaluation of biclustering techniques. *BMC Bioinformatics*, 18(1):55, 12 2017.

[86] S. Panda. Circadian physiology of metabolism. *Science*, 354(6315):1008–1015, nov 2016.

[87] S. Picelli, K. Björklund, O. R. Faridani, S. Sagasser, G. Winberg, and R. Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, 10:1096–1098, 2013.

[88] M. V. Plikus, C. Vollmers, D. de la Cruz, A. Chaix, R. Ramos, S. Panda, and C.-M. Chuong. Local circadian clock gates cell cycle progression of transient amplifying cells during regenerative hair cycling. *Proceedings of the National Academy of Sciences*, 110(23), may 2013.

[89] E. Poggiogalle, H. Jamshed, and C. M. Peterson. Circadian regulation of glucose, lipid, and energy metabolism in humans. *Metabolism*, 84:11–27, jul 2018.

[90] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 2006.

[91] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[92] S. Qian, O. Golubnitschaja, and X. Zhan. Chronic inflammation: key player and biomarker-set to predict and prevent cancer development and progression based on individualized patient profiles. *EPMA Journal 2019 10:4*, 10(4):365–381, 11 2019.

[93] P. Qiu. Embracing the dropouts in single-cell RNA-seq analysis. *Nature Communications*, 11(1), 12 2020.

[94] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, and G. K. Smyth. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), 2015.

[95] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 11 2010.

[96] V. P. Roche, A. Mohamad-Djafari, P. F. Innominato, A. Karaboué, A. Gorbach, and F. A. Lévi. Thoracic surface temperature rhythms as circadian biomarkers for cancer chronotherapy. *Chronobiology International*, 31:409–420, 4 2014.

[97] S. Roweis. Data for MATLAB Hackers.

[98] D. Rugeles, K. Zhao, C. Gao, M. Dash, and S. Krishnaswamy. Biclustering: An application of Dual Topic Models. *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 453–461, 2017.

[99] A. Sarkar and M. Stephens. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nature Genetics 2021 53:6*, 53(6):770–777, 5 2021.

[100] R. Schefzik, J. Flesch, and A. Goncalves. Fast identification of differential distributions in single-cell RNA-sequencing data with waddR. *Bioinformatics*, 37(19), 2021.

[101] E. S. Schernhammer, F. Laden, F. E. Speizer, W. C. Willett, D. J. Hunter, I. Kawachi, and G. A. Colditz. Rotating night shifts and risk of breast cancer in women participating in the nurses' health study. *JNCI Journal of the National Cancer Institute*, 93(20):1563–1568, oct 2001.

[102] J. Seita and I. L. Weissman. Hematopoietic Stem Cell: Self-renewal versus Differentiation. *Wiley interdisciplinary reviews. Systems biology and medicine*, 2(6):640, 11 2010.

[103] J. Sethuraman. A constructive definition of Dirichlet priors, 1994.

[104] A. A. Shafi and K. E. Knudsen. Cancer and the circadian clock. *Cancer Research*, 79(15):3806–3814, aug 2019.

[105] D. D. Shuboni-Mulligan, G. Breton, D. Smart, M. Gilbert, and T. S. Armstrong. Radiation chronotherapyclinical impact of treatment time-of-day: a systematic review. *Journal of Neuro-Oncology*, 145(3):415–427, 11 2019.

[106] N. Singh, D. Baby, J. Rajguru, P. Patil, S. Thakkannavar, and V. Pujari. Inflammation and Cancer. *Annals of African Medicine*, 18(3):121, 7 2019.

[107] C. Stringari, H. Wang, M. Geyfman, V. Crosignani, V. Kumar, J. S. Takahashi, B. Andersen, and E. Gratton. In vivo single-cell detection of metabolic oscillations in stem cells. *Cell Reports*, 10(1):1–7, jan 2015.

[108] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902, 6 2019.

[109] G. Sulli, M. T. Y. Lam, and S. Panda. Interplay between circadian clock and cancer: New frontiers for cancer treatment. *Trends in Cancer*, 5(8):475–494, aug 2019.

[110] G. Sulli, E. N. Manoogian, P. R. Taub, and S. Panda. Training the circadian clock, clocking the drugs, and drugging the clock to prevent, manage, and treat chronic diseases. *Trends in Pharmacological Sciences*, 39(9):812–827, sep 2018.

[111] V. Svensson, R. Vento-Tormo, and S. A. Teichmann. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 13:599–604, 2018.

[112] J. S. Takahashi. Transcriptional architecture of the mammalian circadian clock. *Nature Reviews Genetics*, 18(3):164–179, dec 2016.

[113] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6:377–382, 2009.

[114] P. D. Thomas, A. Kejariwal, N. Guo, H. Mi, M. J. Campbell, A. Muruganujan, and B. Lazareva-Ulitsky. Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Research*, 34(Suppl_2):W645–W650, 2006.

[115] P. Tognini, M. Samad, K. Kinouchi, Y. Liu, J.-C. Helbling, M.-P. Moisan, K. L. Eckel-Mahan, P. Baldi, and P. Sassone-Corsi. Reshaping circadian metabolism in the suprachiasmatic nucleus and prefrontal cortex by nutritional challenge. *Proceedings of the National Academy of Sciences*, 117:29904–29913, 11 2020.

[116] J. Tong, T. Sun, S. Ma, Y. Zhao, M. Ju, Y. Gao, P. Zhu, P. Tan, R. Fu, A. Zhang, D. Wang, D. Wang, Z. Xiao, J. Zhou, R. Yang, S. J. Loughran, J. Li, A. R. Green, E. H. Bresnick, D. Wang, T. Cheng, L. Zhang, and L. Shi. Hematopoietic Stem Cell Heterogeneity Is Linked to the Initiation and Therapeutic Response of Myeloproliferative Neoplasms. *Cell Stem Cell*, 28(3):502–513, 3 2021.

[117] S. Upasham and S. Prasad. Slock (sensor for circadian clock): passive sweat-based chronobiology tracker. *Lab on a Chip*, 20:1947–1960, 2020.

[118] A. Vandenbon and D. Diez. A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data. *Nature Communications 2020 11:1*, 11(1):1–10, 8 2020.

[119] S. Verstovsek, R. A. Mesa, J. Gotlib, R. S. Levy, V. Gupta, J. F. DiPersio, J. V. Catalano, M. Deininger, C. Miller, R. T. Silver, M. Talpaz, E. F. Winton, J. Jimmie H. Harvey, M. O. Arcasoy, E. Hexner, R. M. Lyons, R. Paquette, A. Raza, K. Vaddi, S. Erickson-Viitanen, I. L. Koumenis, W. Sun, V. Sandor, and H. M. Kantarjian. A Double-Blind Placebo-Controlled Trial of Ruxolitinib for Myelofibrosis. *The New England journal of medicine*, 366(9):799, 3 2012.

[120] C. Vetter, H. S. Dashti, J. M. Lane, S. G. Anderson, E. S. Schernhammer, M. K. Rutter, R. Saxena, and F. A. Scheer. Night shift work, genetic risk, and type 2 diabetes in the UK biobank. *Diabetes Care*, 41(4):762–769, feb 2018.

[121] A. M. Vosko, C. S. Colwell, and A. Y. Avidan. Jet lag syndrome: circadian organization, pathophysiology, and management strategies. *Nature and Science of Sleep*, 2:187, 2010.

[122] A. Wagner, A. Regev, and N. Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology*, 34(11):1145–1160, 2016.

[123] W. H. Walker, J. C. Walton, A. C. DeVries, and R. J. Nelson. Circadian rhythm disruption and mental health. *Translational Psychiatry*, 10(1), jan 2020.

[124] H. Wang, E. V. Spyk, Q. Liu, M. Geyfman, M. Salmans, V. Kumar, A. Ihler, N. Li, J. S. Takahashi, and B. Andersen. Cell rep. 20:1061–1072, 2017.

[125] T. Wang, B. Li, C. E. Nelson, and S. Nabavi. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics*, 20(1):1–16, 1 2019.

[126] Y. Wang and X. Zuo. Cytokines frequently implicated in myeloproliferative neoplasms. *Cytokine*, 1(1):100005, 3 2019.

[127] S. Wen, D. Ma, M. Zhao, L. Xie, Q. Wu, L. Gou, C. Zhu, Y. Fan, H. Wang, and J. Yan. Spatiotemporal single-cell analysis of gene expression in the mouse suprachiasmatic nucleus. *Nature Neuroscience*, 2020.

[128] M. G. Wendeu-Foyet and F. Menegaux. Circadian disruption and prostate cancer risk: An updated review of epidemiological evidences. *Cancer Epidemiology, Biomarkers &amp Prevention*, 26(7):985–991, jul 2017.

[129] M. J. Wichurat. Algorithm AS 241: The Percentage Points of the Normal Distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 37(3):477–484, 1988.

[130] G. Wu, R. C. Anafi, M. E. Hughes, K. Kornacker, and J. B. Hogenesch. MetaCycle: an integrated R package to evaluate periodicity in large scale data. *Bioinformatics*, 32(21):3351–3353, 11 2016.

[131] J. Xie, A. Ma, Y. Zhang, B. Liu, S. Cao, C. Wang, J. Xu, C. Zhang, and Q. Ma. QUBIC2: a novel and robust biclustering algorithm for analyses and interpretation of large-scale RNA-Seq data. *Bioinformatics*, 9 2019.

[132] Y. Ye, Y. Xiang, F. M. Ozguc, Y. Kim, C.-J. Liu, P. K. Park, Q. Hu, L. Diao, Y. Lou, C. Lin, A.-Y. Guo, B. Zhou, L. Wang, Z. Chen, J. S. Takahashi, G. B. Mills, S.-H. Yoo, and L. Han. The genomic landscape and pharmacogenomic interactions of clock genes in cancer chronotherapy. *Cell Systems*, 6(3):314–328.e2, mar 2018.

[133] S. H. Yip, P. C. Sham, and J. Wang. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Briefings in Bioinformatics*, 20(4):1583, 3 2019.

[134] E. Yousef, N. Mitwally, N. Noufal, and M. R. Tahir. Shift work and risk of skin cancer: A systematic review and meta-analysis. *Scientific Reports*, 10(1), feb 2020.

[135] Z. Yu, Y. Gong, L. Cui, Y. Hu, Q. Zhou, Z. Chen, Y. Yu, Y. Chen, P. Xu, X. Zhang, C. Guo, and Y. Shi. High-throughput transcriptome and pathogenesis analysis of clinical psoriasis. *Journal of Dermatological Science*, 98(2):109–118, may 2020.

[136] A. Zeisel, A. B. Moz-Manchado, S. Codeluppi, P. Lönnerberg, G. L. Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, and S. Linnarsson. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142, 3 2015.

[137] R. Zhang, N. F. Lahens, H. I. Ballance, M. E. Hughes, and J. B. Hogenesch. A circadian gene expression atlas in mammals: Implications for biology and medicine. *Proceedings of the National Academy of Sciences*, 111(45):16219–16224, 11 2014.

[138] X. Zhang, T. Li, F. Liu, Y. Chen, J. Yao, Z. Li, Y. Huang, and J. Wang. Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Molecular Cell*, 73(1):130–142, 1 2019.

[139] Y. Zhang, X. Xie, P. Wu, and P. Zhu. SIEVE: identifying robust single cell variable genes for single-cell RNA sequencing data. *Blood Science*, 3(2):35–39, 4 2021.

[140] H. Zhao, L. Wu, G. Yan, Y. Chen, M. Zhou, Y. Wu, and Y. Li. Inflammation and tumor progression: signaling pathways and targeted intervention. *Signal Transduction and Targeted Therapy 2021 6:1*, 6(1):1–46, 7 2021.

[141] C. Ziegenhain, B. Vieth, S. Parekh, B. Reinius, A. Guillaumet-Adkins, M. Smets, H. Leonhardt, H. Heyn, I. Hellmann, and W. Enard. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell*, 65(4):631–643, 2 2017.