Transcriptional noise sets fundamental limits to spatially resolved cell-state decoding of the circadian clock.

Anjoom Nikhat[1], Taniya Mandal[1], Nivedha Veerasubramanian[1], Shaon Chakrabarti[1,*]

[1]Simons Centre for the Study of Living Machines, National Centre for Biological Sciences, Bangalore

[*]Email for correspondence: shaon@ncbs.res.in

July 7, 2024

Abstract

Cell-state discovery at single-cell resolution is currently a major endeavor of modern biology. While immense effort has gone into dealing with associated technical noise, there is little appreciation of the fundamental limits imposed by intrinsic biological stochasticity. Using the circadian clock as an example where true cell-states can be precisely defined (the oscillator phase), we study how bursty transcription limits the achievable cellular and spatial resolution of circadian-phase inference. Combining multiplexed smFISH to measure endogenous gene-expression and a novel supervised learning algorithm, we demonstrate how accurate inference is possible using just 3 genes but only after averaging over 20-70 cells. Commonly used algorithms generate single-cell clusters completely different from the true states, successfully recovering them only after averaging. Further, by decoding all states within a population of asynchronized cells, we demonstrate how coarse-graining provides a principled approach to spatially-resolved phase inference. Our results argue that coarse-graining is likely essential for meaningful cell-state assignment across biological systems, even when technical noise is minimized.

Introduction

Rapid advances in multiplexed sequencing and imaging techniques, coupled with parallel developments in theoretical and computational approaches, have opened up the possibility of discovering novel single-cell states across biological systems [1]. These advances are revolutionary since they have the potential to discover heterogeneous cellular states linked to critical functional outcomes, an endeavour that is not possible using traditional bulk measurement techniques [2, 3]. In particular, multiplexed gene-expression measurement technologies have found widespread use in the discovery of cell-states, for example single-cell RNA sequencing (scRNA-seq) [4]. However, these methods typically suffer from large technical sources of noise that hamper true cell-state identification, for example spurious variability arising from PCR amplification biases, differences in cell lysis and reverse transcription efficiency, stochastic molecular sampling during sequencing, ambient RNA etc [5, 6]. As a result, a huge amount of effort has gone into distinguishing true biological variability from measurement artefacts and both experimental and computational normalization methods con-

tinue to be actively developed to overcome these issues [7, 8, 9, 10, 11, 12]. While there has been much progress on understanding the consequences of technical noise, a fundamental question that has received far less attention is whether intrinsically noisy gene-expression may impose limits on cell-state discovery, even in the absence of technical issues. It is well established that transcription of many genes happens in bursts [13, 14], resulting in large mRNA number variability between cells even if they are from the same underlying state. This naturally raises the question of *if* and *when* can robust cell-states be assigned to single cells, and whether the notion of cell-state, traditionally based upon marker genes from bulk studies, can even be applied to a single cell.

To begin addressing this fundamental question, we chose to study the limits of cell-state decoding in the mammalian circadian clock. The clock is ideal to explore this question because of three specific reasons: (1) the circadian clock is an endogenous single-cell oscillator arising from transcriptional and translational feedback loops that combine to generate approximately 24 hour oscillations in a group of 'core' clock mRNAs and proteins [15]. By leveraging the fixed phase relationships between the various clock genes in population-level studies, a well-defined notion of the circadian-state (alternatively the circadian phase) and how to measure it from bulk gene-expression has been established in a series of prior studies [16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28]. (2) The clock mRNAs are expressed in low-copy numbers and demonstrate high intrinsic variability at single-cell resolution [29, 30], thereby making it a good system to investigate the effects of noise on the notion of single-cell states. At the same time however, their expression levels are in fact similar to the vast majority of mammalian genes [31, 32], making the clock fairly representative of expression levels across various biological processes. (3) Finally, decoding clock phase and its heterogeneity within and across tissues is not only a problem of basic biological interest, it is also of significant clinical importance. Emerging evidence indicates that many physiological processes are regulated by the circadian clock, [33, 34, 35], allowing for the exciting possibility of exploiting clock-driven rhythmicity to optimize treatment timing and improve outcomes in various diseases [36, 37, 38, 39, 40, 41]. This however requires knowledge of patient and tissue specific circadian time, and hence inferring the circadian-state from a single tissue sample has of late become an exciting problem in the field of chronobiology [42].

The limits of circadian cell-state decoding from endogenous gene expression have thus far remained unexplored, either in *in-vitro* or *in-vivo*. Historically single-cell circadian clock measurements have

been performed using promoter-driven luciferase reporters in mouse tissue explants [43, 44], or knock-in reporters of proteins [43, 45, 46]. These methods however are limited to model systems due to the necessity of generating genetically modified organisms, and/or do not quantify the endogenous levels of the clock mRNAs. While scRNA-seq has been recently used to measure endogenous clock gene expression, associated technical difficulties necessitate "pseudo-bulk" analyses for accurate circadian phase inference, typically requiring averaging over hundreds to thousands of cells [22, 47]. Though algorithms like Tempo and tauFisher that utilize scRNA-seq datasets improve over older methods like Cyclops [24] and Cyclum [48], they still give large errors in phase inference at the single-cell level. Whether this arises due to intrinsic gene-expression variability or from technical noise invariably associated with scRNA-seq measurements, remains an open question. Crucially, since scRNA-seq loses direct spatial information, it limits our ability to study phase variation across tissue sections. Some studies have utilised single-molecule FISH to obtain more accurate measurements of endogenous mRNA in mouse SCN and liver, but these studies have been limited to a couple of clock genes and do not address the question of limits of circadian phase inference [49, 30].

Here we probe the limits of circadian-state decoding, using accurate measurements of single-cell gene expression in combination with a new phase inference algorithm, PRECISE (**PRE**dicting **CI**rcadian phase from **S**tochastic gene **E**xpression). We utilise a recently developed multiplexed smFISH technique SABER-FISH [50], to obtain absolute mRNA counts of 4 core clock genes simultaneously in hundreds of Dexamethasone synchronised NIH3T3 cells. Using PRECISE we demonstrate that gene expression noise precludes inference of circadian phase from single cells. Applying popular cell-state discovery algorithms using SVD, UMAP, t-SNE and SEACells produces single-cell clusters that are totally different from the known true states. Remarkably however, averaging over 20-70 cells allows for PRECISE phase inference with $\sim$1 hour error using just 3 clock genes, and also allows recovering the true circadian states using all the unsupervised algorithms. Finally we demonstrate how this strategy of coarse-graining over a small group of cells can be used to spatially resolve circadian states in an asynchronized population. Our results demonstrate how gene-expression noise prevents a meaningful definition of single-cell states, but that averaging over a small group of cells allows for an emergent and robust notion of the circadian clock phase.

Methods

Cell culture.

NIH3T3 (mouse; ATCC, CRL-1658) and MLG (mouse; ATCC, CCL-206) were grown in DMEM (Genetix, cat. no. CC3004.05L) supplemented with 10% (vol/vol) serum (Himedia, cat. no. RM10432-500ML) and Penicillin - Streptomycin (Gibco, cat. no. 15140122). All cells were incubated at 37°C and 5% CO_2. In all experiments, cells were grown to about $70 - 80\%$ confluency before fixation and imaging.

SABER-FISH for multiplexed mRNA imaging.

To visualize endogenous mRNA of different clock genes that are produced in low quantities, we utilised a recently developed technique Signal Amplification by Exchange Reaction - Fluorescence *in-situ* Hybridisation (SABER-FISH) [50], with all reagent concentrations and reaction times taken from the updated protocols provided in [51]. We performed multiplexed imaging of 4 clock genes – *Tef, Bmal1, Nr1d1* and *Nr1d2* for fixed monolayers of cells over every 4 hours for 48 hours after Dexamethasone treatment. For each gene, about 50 primary hybridization probes were designed to tile the gene, and extended to various lengths using the Primer Exchange Reaction. These extended probes (or concatemers) were hybridized overnight, following which imager sequences attached to fluorophores were hybridized onto the concatemers. The cells were then imaged on a Nikon Ti2-E epifluorescence microscope with a 60x oil objective, using a Hamamatsu sCMOS camera (ORCA FLASH 4.0). Multiple rounds of imaging were performed by using Formamide washes, which removes the secondary imagers but not the primary probes. A cytoplasmic stain CellMask (Invitrogen, cat. no. H32720) was also included in the different imaging rounds to allow segmentation followed by spot counting per single cell. More details of the experimental protocols and standardisation procedures are provided in Section S1.

Image analysis to quantify single-cell gene expression.

Multiplexed SABER-FISH images were processed using custom image analysis pipeline based on ImageJ and Cellpose [52, 53]. We detected and extracted the spot coordinates for the different genes

from maximum Z-projected images using the RS-FISH plugin in ImageJ [54]. Segmentation of the cells was performed based on contrast-enhanced CellMask images using Cellpose. We wrote custom python-based scripts to perform image registration to identify the same cells across the different rounds of images, and assign spot counts to each individual cell. Further details and parameter values used in the various algorithms are included in Section S2.

Bulk gene expression measurements using qPCR.

Quantitative real-time PCR (qRT-PCR) assay was used to obtain bulk gene expression levels for the different clock genes. We utilised Dexamethasone-resetting to synchronise cells, followed by Trizol based extraction of total cellular RNA using the Purelink RNA extraction kit (Invitrogen, cat. no. 12183018A). cDNA conversion of the cellular mRNA was performed using SuperScript III First-Strand Synthesis System (Invitrogen, cat. no. 18080051). qRT-PCR was performed in 10μL final reaction volume, with three replicates per sample, using iTaq Universal Sybr-Green supermix (Biorad, cat. no. 1725121). We performed qRT-PCR for 2 clock genes – *Bmal1* and *Nr1d2*, and the constitutively expressed gene *Gapdh*. $\Delta\Delta$Ct analysis was performed to obtain the clock genes' expression values as fold change with respect to *Gapdh*. Further details related to the experiment and analysis are included in Section S4.

PRECISE: Predicting Circadian phase from Stochastic gene Expression.

We developed a new algorithm, PRECISE (**PRE**dicting **CI**rcadian phase from **S**tochastic gene **E**xpression), to infer the circadian phase of a test sample from gene expression measurements in a supervised manner. The input training data for PRECISE comprises the mean gene expression levels as functions of time post Dexamethasone synchronization, as well as the errors of the means. At its core, PRECISE uses Gaussian Processes (GPs) with a non-stationary kernel to learn the oscillatory patterns of the different genes from the input training data. PRECISE builds on ODeGP [55], an accurate and sensitive oscillation detection algorithm that we previously developed, to learn the posterior means and variances of the time series of each gene. The power of this approach lies in the fact that GPs provide a non-parametric inference framework in function-space, and provide analytic expressions for the posterior mean and variance (for a basic introduction to GPs and its

application in oscillation detection, see [55].) These expressions for the posterior means and variances are subsequently used to construct a likelihood function that allows calculation of the likelihood of any particular phase given gene expression measurements of a test sample. The likelihood curve tends to have multiple peaks since our training data spans 48 hours post Dex-synchronization. Hence the final predicted circadian phase assigned to the sample, ϕ_{pred}, is calculated by taking the circular mean of the phases corresponding to the likelihood peaks. Mathematical details of PRECISE and its implementation are provided in Section S6.

Assignment of 'true' circadian phase to time-points post Dex-resetting.

The time-points post Dex resetting were converted to phase angles using the procedure broadly described below (for details see Figure S5 and Section S6). The true or global phase of the training or test samples can be determined with respect to any one (arbitrarily chosen) clock gene. For instance, a peak of *Bmal1* could be arbitrarily assigned phase 0 and the next peak as phase 2π. For the training samples, we first bootstrapped the data for a particular gene from each time point to generate mean expression and errors of the mean as functions of time, followed by data normalisation as described earlier in Methods and in Section S6. This normalised input was utilised by ODeGP [55] to determine the posterior mean fit to the data, allowing extraction of the normalised expression values at any arbitrary time point, beyond the experimentally measured ones. Normalized expression was then extracted at discrete time points with a given sampling rate from the posterior mean, and provided as input to pyBOAT [56] for determining phase values. pyBOAT is an algorithm that uses wavelet-transforms to determine the phase of the different samples as functions of time [56]. The phases corresponding to the time points present in our experiments were then extracted and used as phase estimates of the experimental time-points. The phase estimates of the 13 time points used as training data are denoted ϕ_i $(i = 1, 2, ..., 13)$ while the phase estimates of the test samples are denoted ϕ_{true}. More details and parameter values of softwares used are provided in Figure S5 and Section S6.

Calculation of deviation of inferred phase from true circadian phase.

Once PRECISE assigns a predicted phase ϕ_{pred} to a test sample, its deviation from the true phase ϕ_{true} was calculated in two steps as follows (see Section S6 for details):

$$\delta = |\phi_{pred} - \phi_{true}|, \tag{1}$$

$$\Delta(hrs) = \begin{cases} \delta * (24/2\pi), & \text{if } \delta \leq \pi \\ |2\pi - \delta| * (24/2\pi), & \text{if } \delta > \pi \end{cases}$$

Thus the deviation in predicted phase, Δ, always lies between 0 to 12 h.

Unsupervised cell-state discovery algorithms.

We checked whether the results of a variety of unsupervised cell-state discovery algorithms are consistent with PRECISE. To this end, we used SVD (Section S7), lower-rank approximation of the data using SVD followed by an autoencoder with a circular bottleneck layer (Section S8), UMAP, t-SNE and SEACells (Section S10) to discover circadian clock states. Details of each method are provided in the Supplementary Information as indicated.

Results

Single-cell endogenous mRNA imaging recapitulates bulk oscillations and defines cell-states of the circadian clock.

Population level studies have demonstrated that the core circadian clock genes oscillate with ∼24 hour time period and with fixed phase relationships between the genes [15]. Therefore the relative levels of averaged clock gene-expression at various times uniquely define circadian cell-states, interchangeably called 'circadian phase' throughout this text (Figure 1A). These states can equivalently be described by their angular position along the high-dimensional ellipse formed in gene expression space (Figure 1A). To generate cells in these 'true' circadian cell-states, we used Dexamethasone (Dex) resetting to synchronize NIH3T3 cells and then captured the cells at a few discrete

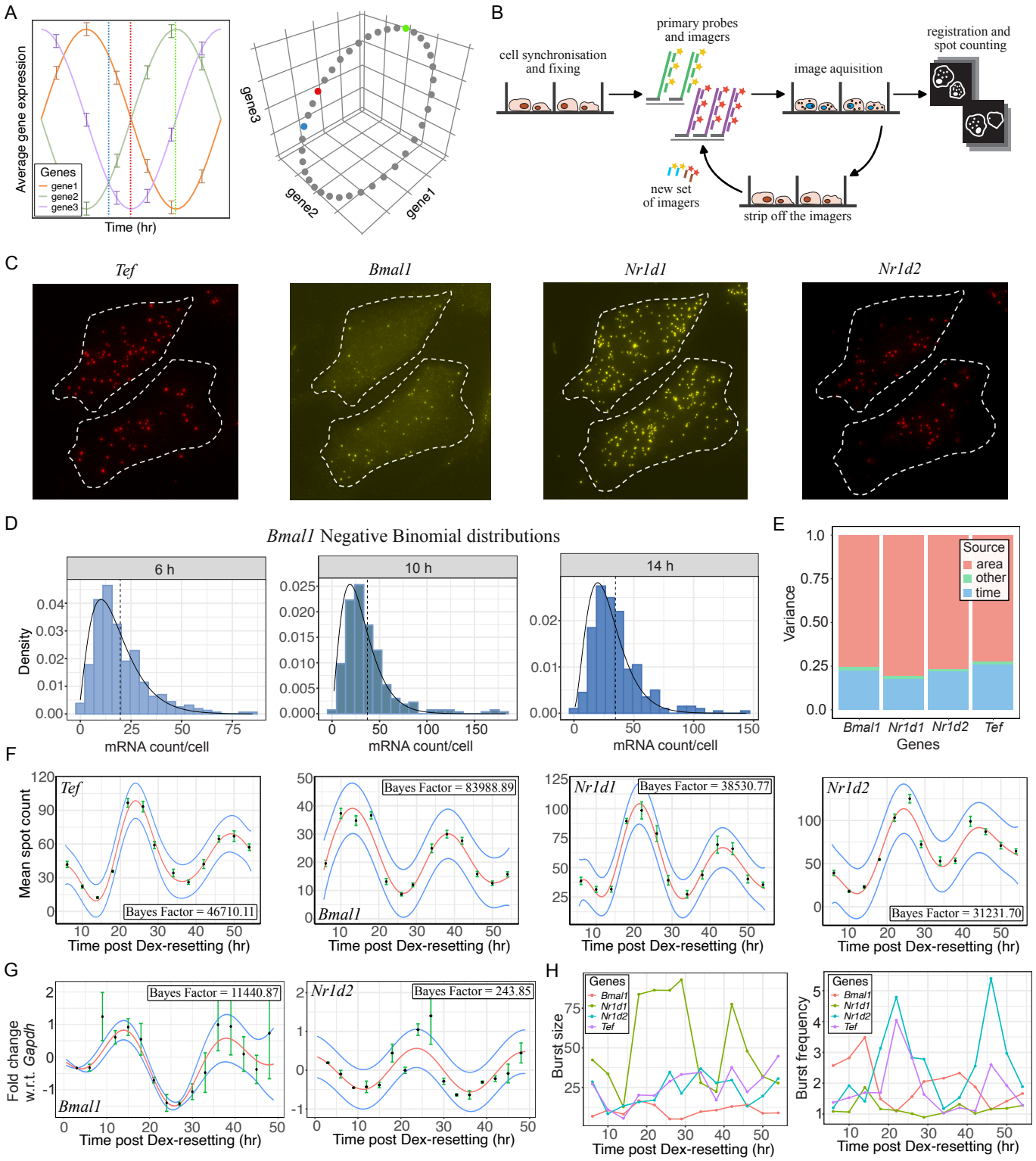

Figure 1: Single-cell gene expression recapitulates bulk properties and defines cell-states of the circadian clock. (A) Schematic showing how the relative levels of average gene expression uniquely defines circadian cell-states. The blue, red and green vertical lines correspond to 3 distinct circadian phases (left) and are positioned at unique angles on the high dimensional ellipse (right). (B) Schematic of the SABER-FISH imaging process. (C) An example of SABER-FISH mRNA spots for the 4 clock genes, shown in the same two cells. (D) Negative Binomial fits to *Bmal1* mRNA counts at three times post Dex resetting, showing a clear shift in the mean. (E) Variance decomposition with area and time as covariates shows a significant contribution from time. (F) Mean mRNA counts and errors in the mean as functions of time. Red lines – ODeGP [55] posterior mean, blue lines – posterior variance. Bayes Factors larger than 14 imply statistically significant oscillations. (G) Bulk qPCR on millions of cells showing the same oscillatory patterns and phase differences for *Bmal1* and *Nr1d2* as observed in panel F. (H) Both burst size and burst frequency show oscillations over time.

times post-Dex for multiplexed gene expression measurements. To reduce technical noise in the measurements as far as possible, we used the recently developed SABER-FISH (Signal Amplification By Exchange Reaction - Fluorescence in situ Hybridisation) method to measure single-cell endogenous mRNA counts [50]. This method uses long repetitive sequences called 'concatemers' at the end of probes complementary to the mRNA sequences of interest, which then act as platforms to hybridise many fluorescently labeled 'imagers' thus amplifying the signal from each mRNA molecule [50, 51]. Besides the signal amplification allowing for sensitive detection of low expression genes, SABER-FISH additionally allows for multiplexing to simultaneously measure multiple genes in the same single cell (Figure 1B). To validate the method, we first imaged a well characterized gene *Cbx5* [50] in the two cell lines used in this study – NIH3T3 and MLG. We consistently achieved $\sim 95\%$ co-localization in both cell types upon using two imagers simultaneously to image *Cbx5* mRNA spots (Figure S1, details in Section S1), thereby demonstrating the accuracy of the measurements.

To collect cells from unique circadian cell-states, we followed a protocol of adding Dex at different times and fixing cells simultaneously to minimize cell density variation amongst different time points (Figure S2, details in Section S1). After fixation, we performed multiplexed SABER-FISH to obtain the absolute mRNA counts of 4 circadian clock genes in $\sim$ 200-250 cells per time point, as shown in Figure 1C (experimental details in Methods and Sections S1 and S2). We first checked that the distributions of mRNA counts for each gene across single cells followed the Negative Binomial distribution (details in Section S3) – a well established signature of transcriptional bursting (Figure 1D shows the distributions for *Bmal1*; other genes are shown in Figure S3). Given the large 'super-Poissonian' noise describing the expression distributions *within* each time-point, we next asked whether expression variation *across* time-points could be detected. We therefore performed variance decomposition analysis [57] (details in Section S3) and found that a significant 20-25% of the total variance could be ascribed to time (Figure 1E), suggesting that average levels of gene expression across time are likely to be well separated despite the high intrinsic noise. To confirm this expectation, we bootstrapped the data from each time point to generate mean expression levels and errors of the mean (Figure 1F). The mean expression for all 4 clock genes showed strong oscillations when analyzed with ODeGP (Figure 1F; details in Section S3), an oscillation detection algorithm that we previously developed and demonstrated to outperform existing methods [55]. Indeed, the oscillatory patterns recapitu-

lated well known phase differences between these four core clock genes, which we also confirmed for a few genes with qPCR (Figure 1G; details in Section S4). Finally, to identify which parameters underlying transcriptional bursting give rise to the expression oscillations, we extracted the burst size and frequency from the mRNA count distributions per time point (see Section S3 for details). We found that both parameters tend to be oscillatory over time (Figure 1H). Therefore, our data suggests that oscillations in both the burst size as well as the burst frequency combine to generate the periodic patterns of gene expression rather than one being constant over time as suggested in previous work [29].

In conclusion, the results of this section confirm that SABER-FISH can be used to accurately measure intrinsic gene expression noise amongst single cells in 4 core circadian clock genes. Recapitulation of strong oscillations in all four genes along with the expected phase differences upon averaging over ~ 100 cells further demonstrated the accuracy of our datasets and measurement technique. Most importantly, these results suggest that the average gene expression levels at different time points post Dex resetting can be used to define unique cell-states of the circadian clock in NIH3T3 cells.

Emergence of gene-expression oscillations with expected phase differences upon averaging over ~ 10 cells.

Given that averages over ~ 100 cells recapitulated known oscillatory patterns and phase differences between the 4 core clock genes, we next explored whether these oscillations and phase relationships could be observed at the single cell level. As observed in Figure 1D, the absolute mRNA counts of the clock genes showed high variance even within a single time-point post Dex synchronization, suggesting that at the single cell level it might be challenging to detect oscillations. We created 'pseudo-time' trajectories by randomly selecting one cell from each time point and plotting the absolute mRNA counts as a function of time (Figure 2A left). We repeated this procedure 50 times to generate 50 such trajectories, depicted by the different colours in Figure 2A (left). If the intrinsic gene expression noise had been low, this procedure should have given rise to similar oscillatory trajectories for all 50 replicates. Instead, the pseudo-time trajectories at the single cell level were characterised by high noise (Figure 2A left and Figure S4) suggesting that placing single cells along an oscillatory trajectory might prove challenging. Remarkably however, when gene expression

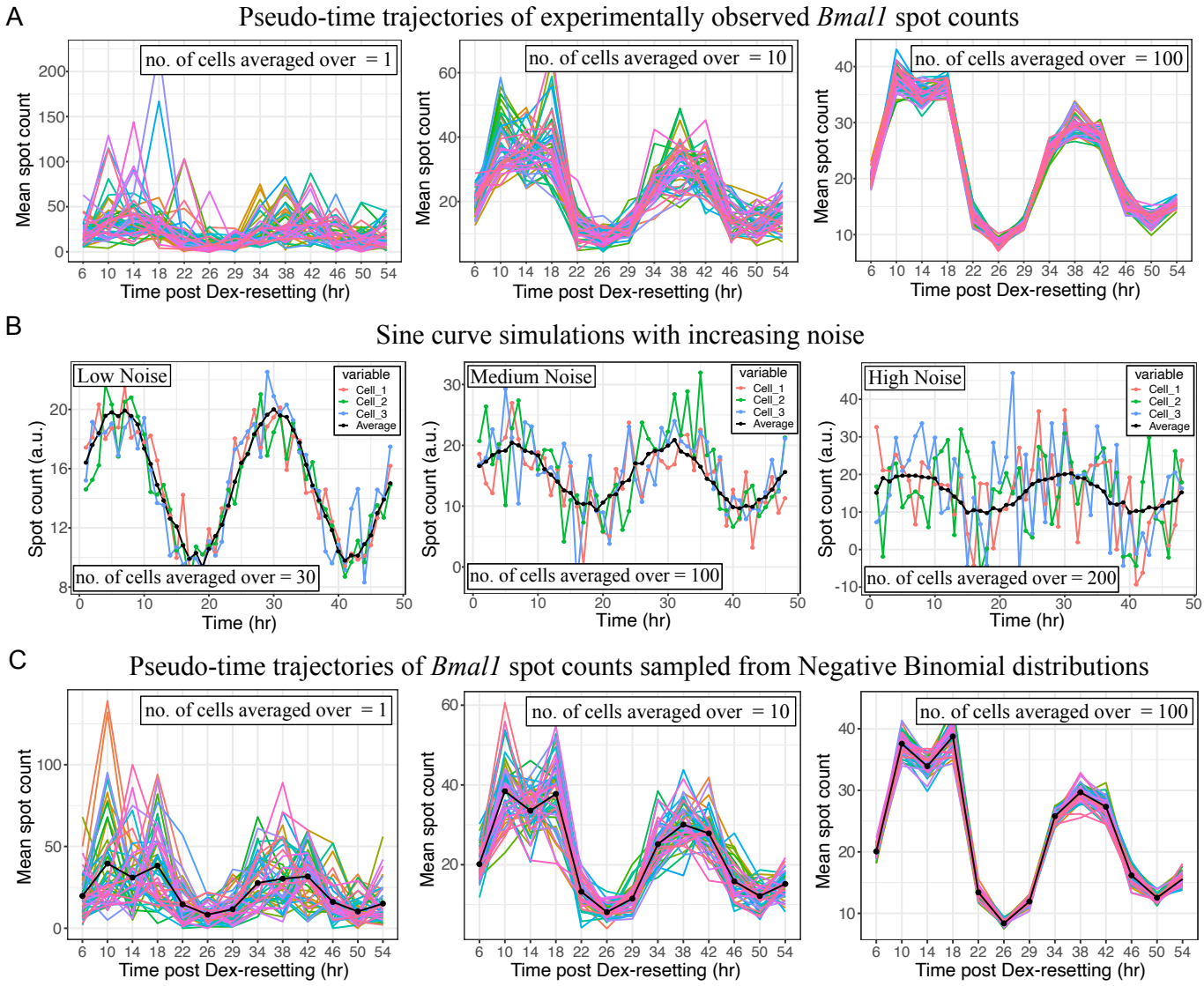

Figure 2: Emergence of oscillations and correct phase differences upon averaging over a small group of single cells. (A) 50 pseudo-time *Bmal1* trajectories of single cells from the experimental dataset (different colours), characterised by high noise (left). Pseudo-time trajectories generated using gene expression values averaged over 10 (centre) or 100 (right) cells demonstrate emergent and low noise oscillations. (B) Simulated sine curves with low (Std. Dev. = 1.5) (left), medium (Std. Dev. = 5)(centre) or high (Std. Dev. = 10) (right) Gaussian noise. Detecting oscillations for the single cells (marked by different colours) is difficult especially in case of high noise. However, oscillatory trajectories emerge upon averaging in all cases (black curves). (C) Pseudo-time trajectories for *Bmal1* generated by sampling from Negative Binomial distributions using parameters obtained from fits to our SABER-FISH datasets from Figure 1D. While single cell trajectories (left) are characterised by high noise, clear oscillations emerge upon averaging over small cellular groups (centre & right), consistent with our experimental observations.

was averaged over groups as small as 10 cells per time point, clean oscillations emerged (Figure 2A centre and right) as well as the correct phase differences between the genes (Figure S4), indicating that assigning unique cell states along a circadian oscillation might be meaningful only after averaging out the transcriptional noise. To gain further intuition on how averaging over cells might lead to reduction in noise and the emergence of clean oscillations as seen in Figure 2A, we carried out simulations using two approaches (details in Section S5): (1) generating time trajectories from a

noisy sine function and averaging over many such trajectories and (2) generating time trajectories by sampling from Negative Binomial distributions with burst size and frequency parameters obtained from Figure 1H, and then averaging over many such trajectories. While approach (1) is purely for demonstrative purposes, approach (2) uses parameter estimates from our smFISH datasets, hence is a good representation of the underlying gene expression noise amongst the clock genes. Both methods demonstrate that even if oscillations are hard to detect in single trajectories (coloured lines in Figure 2B-C), averaging removes this noise and produces clear oscillations (black lines in Figure 2B-C). The more the underlying noise level, the more the number of cells required to observe the emergent oscillations, and hence it is interesting that only ~ 10 cells are sufficient to generate clean oscillations in the experimental dataset (Figure 2A).

In summary, the results of this section suggest that the clock gene expression might to too noisy to robustly assign circadian cell-state or phase to single cells. Intriguingly however, averaging over as few as 10 cells from each time point allowed observation of clear oscillatory patterns along with the expected phase difference between genes, suggesting that averaging over very small groups of cells may be sufficient to overcome intrinsic expression noise and allow accurate assignment of circadian phase.

PRECISE – an accurate supervised learning algorithm for circadian cell-state inference.

Having developed an intuition for the noise levels in clock gene expression and an approximate estimate of the number of cells required to observe emergent oscillations, we then developed a quantitative approach to establish the limits of circadian phase inference in terms of minimum cell and gene numbers required. Since supervised machine-learning approaches tend to outperform unsupervised ones, we first developed a supervised algorithm called PRECISE (**PRE**dicting **CI**rcadian phase from **S**tochastic gene **E**xpression), to accurately infer the circadian phase from mRNA counts of a test sample. PRECISE builds on an algorithm we previously developed using Gaussian Processes (GPs) to learn noisy and non-stationary oscillatory trajectories [55]. Here we provide a brief overview of the method; for more details see Methods and Section S6. The true circadian phases of the training samples are first estimated from the Dex-synchronized time-series of average spot

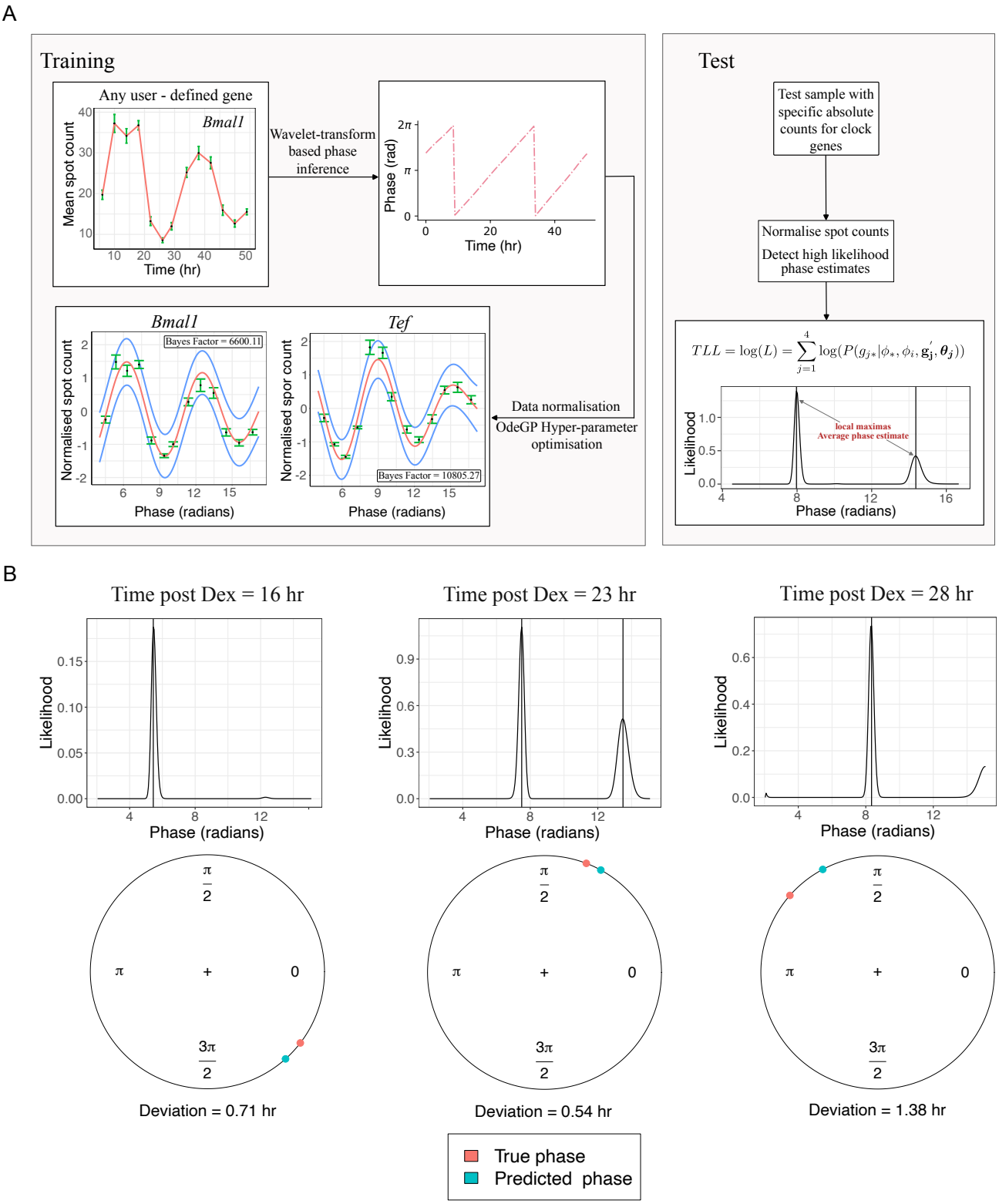

Figure 3: Accurate circadian phase inference using PRECISE. (A) Flowchart indicating the workflow of PRECISE. Training: Mean gene expression at the various times post Dex resetting is first used to assign the circadian phase to the training samples. This is performed using wavelet-transforms. The mean gene expression is then expressed as a function of the circadian phase and normalised. PRECISE then uses Gaussian Process regression to learn the optimised Bayesian posterior mean and variance of the expression levels of the different genes individually. Test: When the raw spot counts of circadian genes for a test sample are provided to pre-trained PRECISE, it normalises the data and constructs the likelihood as a function of the training phases. PRECISE then outputs the predicted circadian phase ϕ_{pred} of the test sample by calculating the circular mean of the phases corresponding to the likelihood peaks. (B) Validating PRECISE for 3 test time points using gene expression data averaged over $\sim$ 200-250 cells and 4 genes. Top: Likelihood versus phase curves show clear peaks, allowing accurate phase estimation. Bottom: quantification of deviations between true and predicted phase for the 3 test cases.

counts, using a wavelet-based approach [56] (Figure 3A and Figure S5). This is followed by normalisation and learning of the optimum Bayesian posterior mean and variance of the expression levels as functions of phase, for each of the four genes individually. Using the posterior mean and variance, PRECISE then constructs the likelihood of a test dataset characterised by a set of clock gene expression values, as a function of circadian phase (Figure 3A). Finally, PRECISE outputs an average phase estimate (ϕ_{pred}) determined by detecting positions of local maxima in the likelihood versus phase curve and obtaining their circular mean.

For validating PRECISE, we collected SABER-FISH spot count data in Dex-synchronised NIH3T3 cells for 5 test time points not included in the training data (8, 16, 23, 28 and 32 hours). For each time point we imaged ∼200-250 cells, and using pyBOAT assigned the true circadian phase to the 3 centre time points 16, 23 and 28 hours (see Methods and Section S6 for why accurate true phase can be assigned only to the centre time points). Using all 4 genes and averages over the entire cell population, we inferred the test sample phase likelihoods using PRECISE (Figure 3B top). We then computed the predicted phases of the test samples from the likelihood peaks (see Methods for details) and compared the predictions against the true phases (Figure 3B bottom). The deviation Δ between true and predicted phases were on average less than 1 hour (Figure 3B bottom), thus verifying that PRECISE can be used to accurately infer circadian phase with only a few genes.

Coarse-graining over 20-50 cells is essential for accurate circadian phase inference using PRECISE.

Having validated the ability of PRECISE to accurately infer circadian phase, we then systematically investigated the effect of decreasing the number of genes and cells on the accuracy of phase inference. We used the same test time points – 16, 23 and 28 hours post Dex-resetting and calculated the deviation of PRECISE-inferred phase from the true phase of the cells (Δ; see Methods and Section S6). We first found that using just 2 genes for phase inference usually produced large values of Δ, though in a few cases combinations of two anti-phasic genes worked well (Figure S6). With 3 genes however, accurate phase inference was always possible, though only after averaging over increasing number of cells. As shown in Figure 4A (23 h test sample is shown; other time points are included in the Figure S7), using only the genes *Bmal1, Nr1d1* and *Tef*, the values of Δ almost uniformly

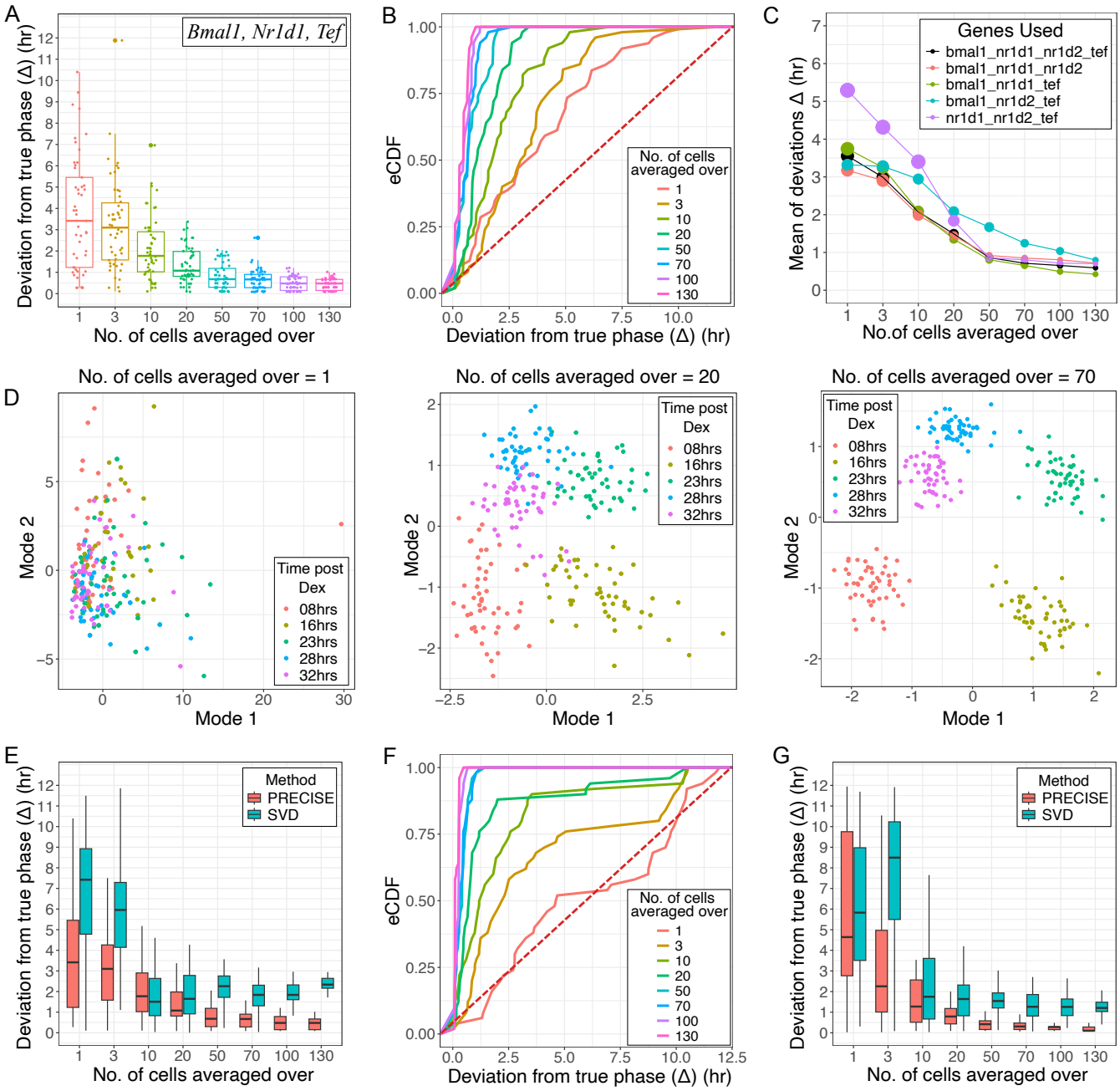

Figure 4: Coarse-graining over small cellular groups is essential for accurate phase inference and recovery of true cell-states. (A) Deviation of PRECISE inferred phase (ϕ_{pred}) from the true phase assigned when genes *Bmal1, Nr1d1* and *Tef* are used. Accurate phase inference is possible when expression values are averaged for small cellular groups > 20 cells as characterised by median $\Delta \sim 1$ hr. (B) Empirical cumulative distribution functions (eCDFs) for Δ. The single cell eCDF is close to the red dotted diagonal line suggesting that phase inference is almost equivalent to random phase assignment. For cell group size ≥ 20, the performance is close to optimum. (C) Comparison of PRECISE's performance for different 3 genes combinations as opposed to using all 4 genes. While all gene sets yielded Δ close to 1 hr, 3 genes allowed phase inference as accurately as 4 genes. Panels A-C correspond to the 23 hr post-Dex synchronisation sample of NIH3T3 cells, where true phase was assigned with respect to *Nr1d1*. (D) Application of SVD to discover circadian cell-states. Left: Projection of single cells in the eigengene space. Cells belonging to different time points (true cell-states) are mixed together with no elliptical structure visible. Centre: Projection of samples when gene expression is averaged over 20 cells shows emergence of pure clusters comprising single time points. Right: after averaging over 70 cells, clear and correct clustering emerges along with the expected elliptical structure, allowing for quantitative phase assignment using the coordinates in eigengene space. (E) Comparison in performance of PRECISE and SVD. The median and variance in Δ are systematically lower for PRECISE for the 23 h sample, demonstrating that PRECISE is significantly more accurate than SVD. (F) eCDFs of Δ when training was performed on NIH3T3 cells but predictions made on MLG cells. (G) Comparison of performance of PRECISE and SVD for the across cell-type phase prediction shown in panel F. Panels F-G correspond to the 16 hr post-Dex synchronisation sample of the MLG cell line. True phase here was assigned with respect to *Tef*.

covered the full range of 12 hours when single cell gene expression was used. This demonstrates that circadian phase cannot be robustly decoded from single cells. However, upon averaging gene expression over as small a group as 20 cells, PRECISE was able to predict the circadian phase with a median $\Delta \sim 1$ h (Figure 4A), which is comparable to errors generated by supervised algorithms like ZeitZeiger using gene expression data averaged over millions of cells and at least 10-13 genes [16]. Δ went down further to ~ 0.5 hours upon averaging over 130 cells, and the variation in Δ shrunk to less than an hour, suggesting that PRECISE is an extremely accurate phase decoder (Figure 4A). We also visualized these deviations of inferred phase using empirical cumulative distribution function (eCDFs), to better understand the performance of PRECISE. The eCDF using single cells was close to the diagonal line (red dotted line in Figure 4B), suggesting that predictions using single cells are only slightly better than making purely random choices of phase. For averages over ≥ 20 cells, the eCDFs were close to optimum with minimum performance improvement for higher sample sizes, further emphasizing that gene expression when averaged over small cellular groups allows for accurate phase estimation (Figure 4B). Finally we compared Δ when different combinations of 3 genes were used as well as when all 4 genes were used (Figure 4C). While all gene sets yielded Δ close to 1 h, interestingly 3 clock genes allowed phase inference as accurately as 4 genes (Figure 4C), suggesting that the phase information provided by gene expression likely saturates at only 3 genes. All these results were consistent irrespective of which gene was used to assign true phase and the time point analyzed (Figure S8). The results in this section thus far are broadly consistent with the intuition developed earlier, demonstrating that noisy circadian clock gene expression precludes accurate circadian phase inference from single cells, but averaging expression levels over small cellular groups of 20-50 cells allows for accurate phase assignment.

Like any supervised algorithm, PRECISE makes assumptions regarding the underlying model, noise structure, and additionally requires time-course gene expression data with wavelet based 'true-phase' assignments. To confirm that these modeling choices do not qualitatively affect our results, we next explored the performance of unsupervised learning algorithms that do not demand the pre-requisites mentioned. Early work demonstrated how Singular Value Decomposition (SVD) can be used to reduce the dimensionality of complex oscillatory gene expression datasets and identify sets of "characteristic modes" or "eigengenes" that represent the global expression profiles [58]. Projection

of samples onto the eigengene space leads to emergence of elliptical structures where the angular positions correspond to specific phases of the oscillatory cycle. In order to apply SVD to our test dataset we first generated a gene $\times$ cell matrix of dimensions 3×250 (taking the same *Bmal1, Nr1d1* and *Tef* combination as shown in Figure 4A-B), whose entries were expression values of 50 randomly chosen single cells from each of the 5 different time points. As shown in Figure 4D (left), projection of the single cell data onto the top 2 eigengene space failed to cluster cells from the same time point together. We then generated data matrices whose entries were no longer single cell expression values, but averages over 20 (Figure 4D centre) or 70 (Figure 4D right) cells. 50 such average values per time point were generated by bootstrapping to create matrices of dimensions 3×250, and SVD was performed as before. Consistent with the results of PRECISE, we observed that even with 20-cell averages the emergent clusters comprised cells from specific time-points only, and a clear elliptical structure emerged with the correct ordering of the time points at the 70 cell average level (Figure 4D right). We noticed that the samples from the 8 and 32 h time points do not cluster together as might be anticipated. This happens because the circadian clock's period is not exactly 24 hours, hence they do not represent the same phase. These results clearly demonstrate that unique cell states can be defined only after averaging over 20 cells or more.

In order to quantitatively compare the performance of SVD relative to PRECISE we inferred the circadian phase of the individual samples by obtaining the tan^{-1} of the ratio of their coordinates in the eigengene space and subsequently calculating Δ. We observed that the median and variance of Δ for almost all cell averages were lower for PRECISE in comparison to SVD, suggesting more accurate phase estimation using our supervised learning algorithm (Figure 4E) (details in Figure S9 and Section S7). Additionally, since previous work has demonstrated the superiority of autoencoders in cyclic phase reconstruction, we checked that an autoencoder applied to the results of SVD makes no difference to our results (details in Section S8 and Figure S10).

Finally, we also explored the possibility of across-cell line phase predictions by performing inference on mouse lung fibroblasts (MLG), when PRECISE was trained using time course data from NIH3T3 cells (details in Section S9). For the 16 h post-Dexamethasone sample, the eCDFs of Δ for single cells was close to the diagonal line (red dotted line in Figure 4F). However, when the expression was averaged over cell numbers ≥ 50 cells, we obtained Δ close to 1 h (Figure 4F-G) suggesting accurate

across cell-lines phase prediction. When we performed SVD-based phase inference and compared it with PRECISE's estimates, we once again observed better performance in case of PRECISE (Figure 4G). Results from the other time points and gene combinations are shown in Figures S11-S12.

In summary, both the supervised and unsupervised algorithms demonstrate that the true clock phases cannot be recovered at the single cell level due to the large intrinsic noise from bursty transcription. However, accurate inference with just ~ 1 hour deviation is possible with only 3 genes after averaging over a small group of cells (20-50 cells for PRECISE, ~ 70 for SVD). Interestingly, accurate phase inference can also be achieved across cell-lines by training on one cell type and predicting on another, suggesting that the phase differences between the core clock genes are broadly consistent between tissues.

Transcriptional noise prevents meaningful clustering of single cells by cell-state discovery algorithms.

We next asked if either of the commonly used non-linear dimensionality reduction techniques, UMAP (Uniform Manifold Approximation and Projection) [59] or t-SNE (t-distributed stochastic neighbor embedding) [60] allow for better clustering of single cells according to their known true circadian cell-states. Note that unlike PRECISE or SVD, these non-linear methods are not expected to allow for quantitative inference of phase since the elliptical structure (Figure 1A) will not be maintained after projection into lower dimensions. We used both UMAP and t-SNE on our absolute mRNA count data and performed k-means clustering on euclidean distances between the projected points to obtain 5 clusters in each case (Figure 5A top; details in Methods and Section S10). Annotating the cells based on their known true circadian phases demonstrated that cells from different circadian phases were grouped into the same clusters (Figure 5A bottom), further emphasizing the fact that inherently noisy circadian gene expression limits correct circadian phase assignment at single cell resolution.

Besides the widely used UMAP and t-SNE algorithms for cell-state discovery, recent efforts have gone towards partitioning scRNA-seq datasets to identify 'metacells' representing distinct cell-states, where the variability within a metacell arises from technical rather than biological sources [61, 62]. We implemented a recently developed algorithm, SEACells, to infer 5 metacells in our dataset

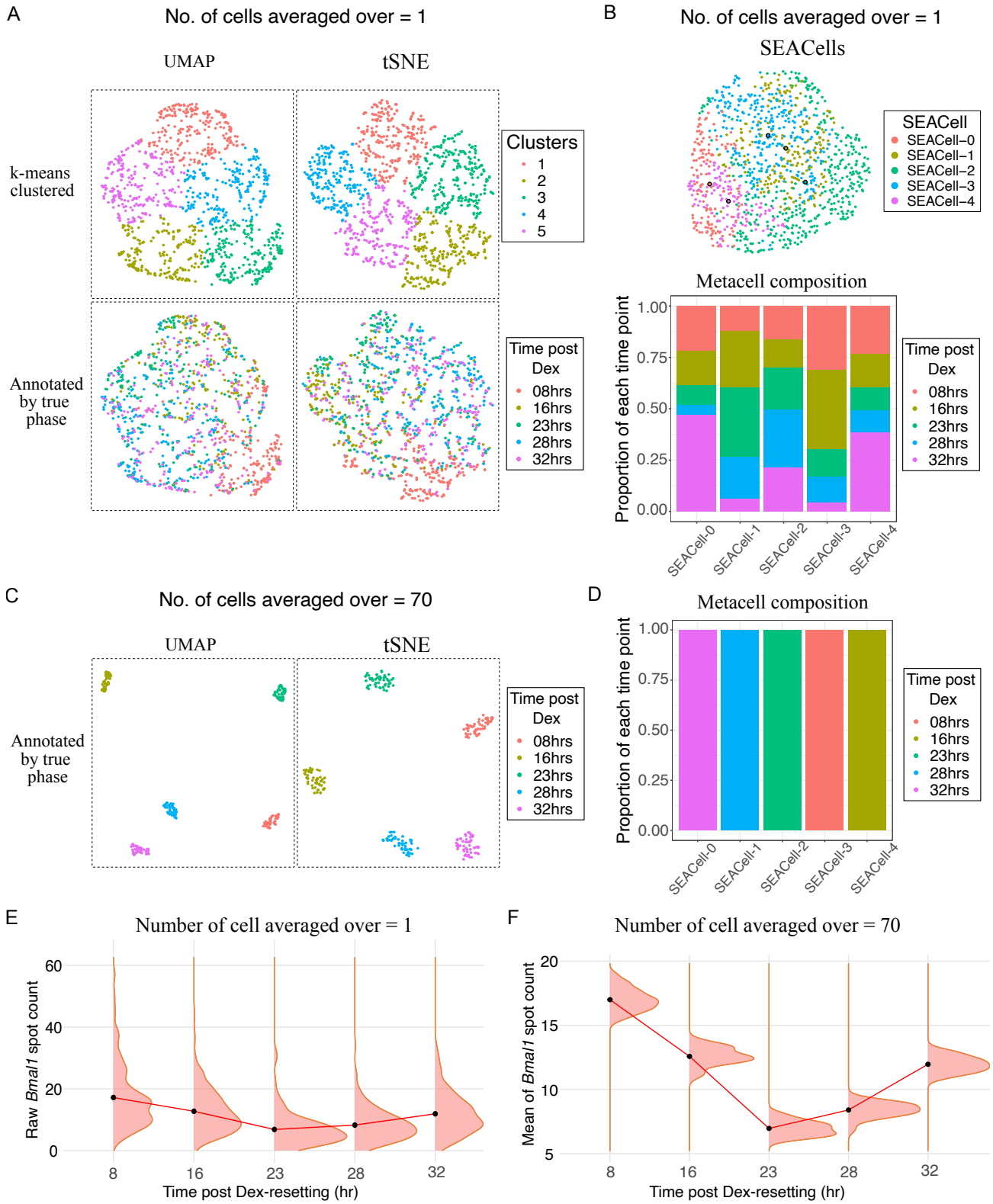

Figure 5: Transcriptional noise prevents meaningful clustering of single cells by cell-state discovery algorithms. (A) Top: 5 single cell clusters (cell-states) identified using UMAP and t-SNE dimensionality reduction followed by k-means clustering on the euclidean distances. All four genes *Bmal1, Nr1d1, Nr1d2* and *Tef* were used in this process. Bottom: cells annotated by the known true cell-states (times post Dex resetting) in the UMAP and t-SNE projections. Comparison with the top panels show that each of the UMAP and t-SNE clusters have a mix of cells from various true states. (B) Top: SEACells based identification of 'metacells' (dark circles) for our single cell circadian clock dataset. Bottom: Metacell composition shows the proportion of cells from different true circadian-states that are included in each metacell. Similar to panel A, each SEACell identified metacell comprises cells from different true states. (C-D) Application of UMAP, t-SNE and SEACells on a dataset averaged over 70 cells successfully identifies pure clusters, with each cluster comprising cells from only one particular true circadian state.

Figure 5: (E) Distributions of raw spot counts for the gene *Bmal1* in single cells across the 5 test time points post Dex-resetting. While the means of the distributions (black dots) were shifted, the distributions were highly overlapping, thereby limiting cell-state inference at single-cell resolution. (F) Distributions of mean *Bmal1* spot counts across time post Dex, when gene expression was averaged over groups of 70 cells. The distributions of the means are clearly well separated, allowing accurate cell-state discovery.

and checked what proportion of cells from different time points were classified into a single metacell [62]. With accurate classification, cells from a single time point would be expected to dominate the population of cells included in a single metacell. However as shown in Figure 5B, all the metacells identified encompassed cells from different time points, suggesting that like UMAP and t-SNE, SEACells also fails to identify the true circadian cell-states as a consequence of the underlying transcriptional noise. Furthermore, the identified cell-states were not robust – the 3 methods grouped very different sets of single cells into clusters, which further depended on the specific parameter choices (Figure S13, Methods and Section S10).

Since the results in the previous section had demonstrated our ability to accomplish accurate phase inference only after averaging, we next tried the same strategy on all the 3 methods. We generated 50 samples per time point, where each sample was obtained by averaging gene expression over 70 randomly chosen cells from within that time point. We put together all these samples from 5 time points and explored the performance of each of these algorithms on the averaged dataset. Consistent with the results of our last section, we now observed that samples from the same time point were indeed clustered together after performing k-means clustering on both the UMAP and t-SNE projections (Figure 5C). The same was true for SEACells – each metacell identified now comprised samples from only single time points, as shown in Figure 5D. Therefore irrespective of the cell-state discovery algorithm we used, we could recover the true circadian states only after averaging.

To summarize, the results in this section demonstrate that irrespective of the algorithm used, meaningful and robust circadian cell-states cannot be assigned at the single-cell level, but require some level of coarse-graining before they can be discovered in an unsupervised fashion. These results may have significant implications for the discovery of cell-states in the broader context of other noisy biological processes.

Biological, not sampling noise prevents correct single-cell state assignments.

To understand what enabled accurate cell-state assignment for cell groups but not single cells, we visualized the distribution of raw spot counts for the clock genes across the various time points (the distributions for *Bmal1* are shown in Figure 5E, the rest are shown in Figure S14). While the means of the distributions (black dots in Figure 5E) were shifted as expected from our analysis in Figure 1F, the distributions were strongly overlapping for the different circadian states. However, the distribution of the means were well separated since the variance of each distribution was much reduced (Figure 5F). Note that though the means of the 16 and 32 h time points are the same (Figure 5F), the presence of the other genes will allow for unique state assignments. All the cell-state discovery algorithms, with their inherent approach of clustering cells with statistically similar expression profiles, could therefore discover the correct states only after averaging.

Could sampling noise have contributed to the non-identifiability of single-cell circadian states? In sequencing based methods like scRNA-seq, only a small fraction of the total cellular mRNA (upto $\sim 30\%$ with modern protocols [63]) can be captured, leading to sampling artefacts. However, single-molecule imaging based methods like smFISH largely overcome this issue. In our SABER-FISH setup, the high colocalization fraction of 90-95% (see Figure S1 and Section S1) combined with simple probability arguments demonstrate that our probes capture around 95% of the total transcripts (see Section S11 for details). To further check if the 5% of missed transcripts could potentially lead to issues of cell-state identification, we ran a number of simulations. We generated synthetic distributions of single-cell mRNA counts for four different genes, representing two distinct circadian states or phases. We utilized the NB parameters derived from fits to our training datasets, resulting in the underlying distributions being strongly overlapping. For each of the two states, we sampled 200 or 1000 individual cells each characterised by a vector of absolute counts of 4 genes sampled from the underlying distribution. The counts per cell thus generated are the 'true counts' or the actual mRNA numbers in different cells. From these, we ultimately obtained the 'observed counts' by sampling from the true counts with some probability <1. We applied UMAP to the observed counts matrix of mRNA per cell and annotated the samples by their true phases to analyze their clustering patterns. As shown in Figure S15, regardless of whether the sampling probability was

low (0.20-0.30; characteristic of scRNA-seq) or high (0.90-0.95; characteristic of smFISH), when the underlying distributions were overlapping (as is the case with the circadian clock), state inference was possible only after averaging.

Overall, this section demonstrates that our experimental observations on the circadian clock states (correct clusters emerge only upon averaging) arises when biological noise drives the gene expression distributions of different states to strongly overlap with one another. In such cases, irrespective of the level of sampling noise, single-cell state assignments will not reflect the true circadian states, and averaging becomes essential to discover the true states.

Averaging over locally proximal cells allows spatially-resolved circadian phase inference.

In the last sections we demonstrated that single-cell circadian state inference is fundamentally limited by the transcriptional noise associated with clock gene expression. However, averaging over small cellular groups from within the same state leads to reduction of this noise, thus allowing accurate phase estimation for synchronised populations of cells. A key question that therefore arises is regarding the selection of cells to be averaged over, in order to accurately represent the underlying cell-state. For a synchronized population of cells where the underlying assumption is that all cells are in the same state, a random choice of cells is clearly appropriate. When we performed phase inference for 50 samples, each of which was derived from gene expression averaged over 20 randomly chosen cells from our Dex-synchronised test dataset, we observed that the inferred phases were localised to specific regions on the circle (Figure 6A). To quantitatively measure the phase coherence, we computed the Rayleigh statistic, whose value is close to 1 when the data is localised on a circle and 0 when the data is uniformly distributed (see details in Section S12). The high Rayleigh statistics observed for the synchronised population of cells (Figure 6A) as well as the clear separation of clusters seen earlier (Figures 5C-D and 4D) demonstrate that random selection of cells is sufficient.

On the other hand, how should one select cells for averaging from an asynchronized population? This is typically the case for most biological systems (for example developmental cell-states), which unlike the circadian clock cannot be well synchronized or are not often found in synchronized states.

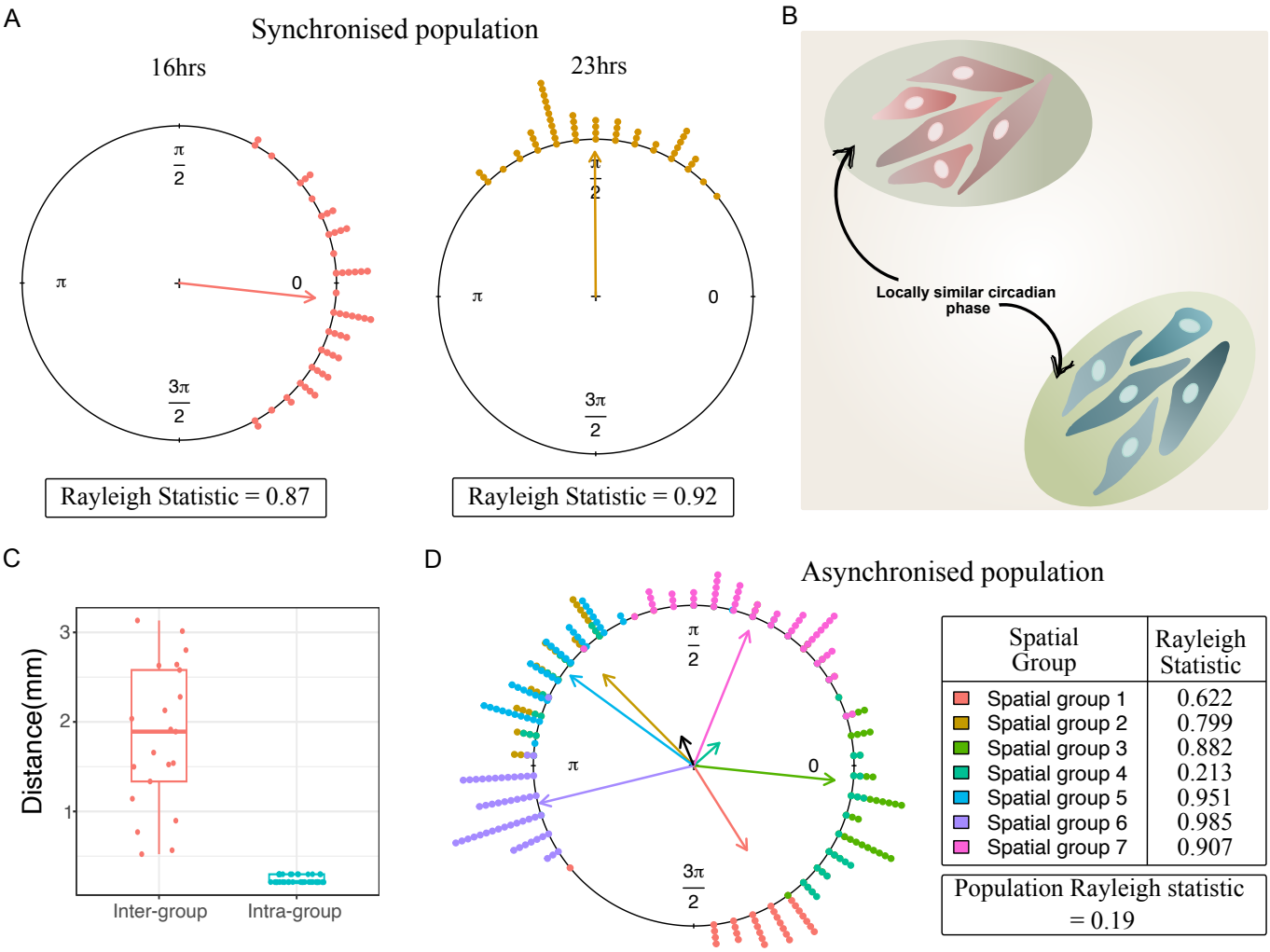

Figure 6: Averaging over locally proximal cells allows spatially-resolved circadian phase inference. (A) In synchronized populations, circadian phase was assigned 50 times (solid circles) via averaging over 20 randomly selected cells on each occasion (left-16 h and right-23 h time points). The Rayleigh statistic quantifying the spread in the inferred phases is close to 1 in both cases, suggesting strong phase coherence. (B) For an asynchronous population, local regions of cells are expected to have similar circadian phases due to paracrine signaling or lineage effects, but distinct phases will be maintained for spatially distant cell groups. (C) Comparison of the distances within and across 7 spatial groups of cells that were imaged in an asynchronous population of NIH3T3 cells. The distances of cells within each group (intra) were distinctly less than the distances across groups (inter). (D) Circadian phases inferred for different spatial groups of cells (denoted by different colours) in the asynchronized population. Similar to panel A, each spatial group comprised 50 phase measurements (solid circles) after averaging over 20 cells randomly chosen from within that group. The overall Rayleigh statistic is close to 0 (black arrow), demonstrating a near-uniform phase distribution across space. However, high phase coherence is observed within the different spatial groups as quantified by the high individual Rayleigh statistics (coloured arrows).

For such asynchronous populations comprising cells from various states, spatially proximal cells are likely to be representative of the same cell-state. In the context of the circadian clock, a recent study suggested that paracrine signaling via TGF-β causes local regions of cells to have similar circadian phases but spatially distant cells to have distinct phases (Figure 6B) [64]. Furthermore, if phases are decoded correctly, a uniform circadian phase distribution is expected to emerge from an asynchronous population, as has previously been demonstrated using live-cell imaging [65]. We utilized these two

ideas to see if spatially resolved phase inference could be performed reliably. We carried out SABER-FISH on an asynchronous population of NIH3T3 cells (details in Section S12), followed by phase inference using PRECISE on spatially averaged gene expression values. In total we had 7 separate spatial groups in our data, and we confirmed that these groups were sufficiently separated from each other as indicated by larger inter-group distances compared to the intra-group distances (Figure 6C and Figure S16). For the cells belonging to a particular spatial group, we performed 50 instances of phase prediction, where for each instance gene expression averaged over 20 randomly selected cells was used as input. We executed similar phase inference for each of the remaining spatial groups present in our data and visualised the inferred average phase (ϕ_{pred}) on a circle. As shown in Figures 6D and S17, both the expectations were met – (i) a uniform distribution of inferred phases was found for the entire population (black arrow corresponding to the population Rayleigh statistic) and (ii) high phase coherence was observed within each spatial group as characterised by the high individual Rayleigh statistics (colored arrows). Together, these results suggest that choosing spatially proximal cells may be good strategy to perform the necessary averaging of gene-expression before decoding circadian states.

To summarize, averaging over random groups of cells is sufficient to detect the true circadian cell-states only in synchronized cell populations. However, when dealing with an asynchronized population that is expected to have cells in a mixture of states, we suggest a simple strategy of averaging over spatially proximal cells to recover the true underlying states. Our results provide a principled approach towards decoding circadian phase in a spatially resolved manner.

Discussion

Spatial mapping of single-cell states, particularly using gene expression measurements, is an exciting emerging frontier across various fields of biology. However, fundamental limits to this problem imposed by intrinsically noisy single-cell gene expression have rarely been systematically studied. Indeed, an implicit idea that seems to pervade many fields of single-cell biology is that the accurate assignment of cell states from gene expression is ultimately limited by technical issues – that overcoming measurement noise (either by developing more precise experimental techniques or better computational de-noising algorithms) would eventually allow for meaningful identification of single-

cell states [2]. In this work, we use the specific example of the mammalian circadian clock to explore how bursty gene expression in single cells sets limits on decoding cell-states, and how these limits in turn affect the resolution at which spatial state-inference is achievable. Using accurate smFISH-based gene expression measurements to minimize technical noise and a variety of phase inference algorithms, we demonstrate that the circadian cell-states precisely defined using population-averaged gene expression from synchronized cells cannot be recovered in single cells (Figure 4). Indeed, none of the popular cell-state discovery algorithms could identify meaningful circadian states, generating clusters comprising mixtures of cells from every true circadian state (Figure 5). Remarkably however, averaging over as few as 20-70 cells allows every algorithm to robustly recover the true circadian cell-states, a cell number that is sufficiently small to allow for high spatial resolution but large enough to average out the biological noise. Our results therefore caution against assigning circadian phase to single cells using gene expression measurements, but suggest that averaging over small groups of cells *before* performing phase inference or dimensionality reduction would allow for assignment of meaningful circadian phases.

Are there any parallels to our results that can be found elsewhere? We found a couple of earlier works that are conceptually related to our findings [66, 31]. In the more recent study, Squair and colleagues demonstrated that computing differential expression at single-cell level invariably leads to false discoveries as a result of biological variability, and that this issue is circumvented after 'pseudo-bulking' [66]. While naively this result may seem unrelated to our study which focuses on cell-states, ultimately the very notion of a cell-state inherently depends on relative expression levels of various genes (Figure 1A). As a result, fundamental limitations in discovering differentially expressed genes in single cells are bound to confound discovery of cell-states, as our work demonstrates. The second and older study used scRNA-seq with careful spike-in controls to distinguish biological versus technical noise, and demonstrated that to recover the same transcriptome and number of measurable genes as obtained from bulk RNA-seq measurements, groups of 30-100 cells need to be averaged over [31]. This number is remarkably similar to our finding of 20-70 cells, though our results are derived from a very different system. These results together suggest a general principle – even if technical noise can be minimized, averaging over a small group of cells as opposed to using single cells is likely to allow more robust and meaningful assignment of cell-states in a variety of biological contexts.

It is crucial to note that here we *begin* with a clear notion of cell states (the true circadian states) based on measurements on synchronized cells (Figure 1A). This ultimately allows us to compare single-cell state assignments by various algorithms with the known true states. This comparison is usually not possible in the typical cell-state discovery process followed in the literature, where *a priori* definitions of cell states are not available. These methods typically cluster cells by similarity in gene expression, hence in our case where distributions of each state overlap significantly due to transcriptional noise (Figure 5E), they end up clustering single cells from different true circadian states (Figure 5A-B). However, are these single-cell clusters meaningful in any interpretable way? If the cell-states of interest have highly differentially expressed genes driving differences between the states, single-cell clusters are likely to be meaningful. A classic example is the cell-cycle, where G1, S, G2 and M phase assignments can be accurately made for single cells [67, 68]. Similarly, cell-type assignment based on a few highly expressed marker genes is also usually possible [69] (though this can also pose challenges if marker genes are not highly expressed [70]). However, given that the vast majority of mammalian genes are expressed at only 1-50 transcripts per cell [31] (similar to the circadian clock genes) with a median of just 17 transcripts [32], it is likely that most cell-states will be governed by subtle differences in mRNA numbers and hence their identification strongly affected by expression noise. In our circadian clock example, this argument is supported by our observation of non-robustness of the states identified at single cell level – the cluster compositions varied not only between the different algorithms used, but also between different parameter choices of a given algorithm. In contrast, all the unsupervised algorithms, SVD (with and without an additional autoencoder), UMAP, t-SNE as well as SEACells robustly discovered exactly the same true states *after* gene expression was averaged over just 20-70 cells (Figure 5C-D). In the very least, our results demonstrate that for biology driven by low expression genes, single cell clusters will very likely be completely different from those discovered after averaging over a few cells. This naturally poses a question that one has to confront – which definition of cell-state, the coarse-grained one or the one at single-cell level, is more biologically relevant.

If one argues that the coarse-grained definition of cell-state is more meaningful, the natural question that arises is how to decide which group of cells to average over? As we demonstrated, the averaging needs to be done within cells from the same state (Figure 5F). However, in most cases one does

not have synchronized populations and hence no notion of the cellular states to begin with, making this a challenging problem to address. We provide one working solution – to assume that spatially proximal cells belong to the same state due to paracrine signaling or lineage relationships, and hence averaging the expression levels of cells in very small spatial neighborhoods will allow recovery of the true clusters. We demonstrated the validity of this approach by reproducing the expected uniform distribution of circadian phases in an asynchronized cellular population, with stronger phase coherence within local spatial neighborhoods validating the assumption of local similarity of cell states (Figure 6D). When no spatial information is directly available as with typical scRNA-seq datasets, one could potentially use 'landmark' genes [71] to get an indirect sense of spatial neighborhoods before averaging. However, rigorous solutions to this problem remain an open and exciting challenge. Recent approaches based on careful *a priori* definitions of expression distributions underlying cell states [12] may provide ways of addressing this question, and remain to be explored in future work.

Our study has a few limitations that will need to be clarified in future work. Main amongst these is the fact that all our results were derived from cell lines. While transcriptional bursts have been reported in tissues [72], there might be additional noise reduction mechanisms in the *in vivo* context that may change the level of coarse-graining suggested by our current results. Additionally, it could be argued that Dex resetting is unlikely to be perfect in synchronizing the single-cell circadian clocks, leading to a mixture of true states even within a single time-point post Dex addition. While possibly existent to some degree, this limitation is unlikely to affect our results significantly. Since Dex is a well established strong resetting agent generating 'type-0' phase-transition curves, most single cells are expected to be in the same state upon Dex treatment [65]. If there were a number of other states mixed in due to incomplete Dex action, one would have seen one large cluster along with some small clusters upon phase inference. We do not see this behavior – single cell phase inferences are uniformly distributed (Figure 4A) and clusters are also well mixed in the true states (Figure 5A-B), ruling out any significant effects of incomplete Dex synchronization.

Conclusion

Measuring the internal circadian time or phase of cells and tissues from single gene-expression snapshots is a major endeavor in chronobiology, with important fundamental as well as clinical implications. We demonstrate that due to transcriptional noise, circadian phase can accurately be decoded only after averaging expression levels of 3 core clock genes over small groups of $\sim$ 20-70 cells – small enough to allow high spatial resolution in phase inference, but large enough to overcome transcriptional noise. We explicitly demonstrate how a unique and robust notion of the circadian cell-state emerges only after averaging, suggesting that single-cell circadian phase assignments are unlikely to be biologically meaningful. Since the vast majority of mammalian genes are expressed at levels similar to that of the circadian clock, our results raise broader and fundamental questions regarding the essentiality of coarse-graining in defining cell-states in other noisy biological processes as well.

Author contributions

S.C. conceptualized and designed the study with help from A.N., N.V. established and optimized the SABER-FISH protocol, T.M. developed the multiplexed version of SABER-FISH, collected training datasets with A.N. and performed initial data analysis, N.V. and T.M developed the initial image analysis pipelines, A.N. collected all test datasets across cell lines, built the final image analysis and registration pipelines, developed the supervised algorithm PRECISE and performed the computational analyses, A.N. and S.C. wrote the paper with inputs from N.V. and T.M., S.C. supervised the work and raised funding for the project.

Acknowledgments

S.C. acknowledges funding from SERB (Government of India) under project number SPR/2021/000486 as well as intramural funds from National Center for Biological Sciences–Tata Institute of Fundamental Research (NCBS-TIFR).

References

[1] Alev Baysoy et al. "The technological landscape and applications of single-cell multi-omics". In: *Nature Reviews Molecular Cell Biology* 24.10 (Oct. 2023). Publisher: Nature Publishing Group, pp. 695–713.

[2] Cole Trapnell. "Defining cell types and states with single-cell genomics". In: *Genome Research* 25.10 (Oct. 2015). Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, pp. 1491–1498.

[3] Samantha A Morris. "The evolving concept of cell identity in the single cell era". In: *Development* 146.12 (2019), dev169748.

[4] Aleksandra A. Kolodziejczyk et al. "The Technology and Biology of Single-Cell RNA Sequencing". In: *Molecular Cell* 58.4 (May 2015), pp. 610–620.

[5] Catalina A. Vallejos et al. "Normalizing single-cell RNA sequencing data: challenges and opportunities". In: *Nature Methods* 14.6 (June 2017). Publisher: Nature Publishing Group, pp. 565–571.

[6] Lukas Heumos et al. "Best practices for single-cell analysis across modalities". In: *Nature Reviews Genetics* 24.8 (2023), pp. 550–572.

[7] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome biology* 15 (2014), pp. 1–21.

[8] Catalina A Vallejos, John C Marioni, and Sylvia Richardson. "BASiCS: Bayesian analysis of single-cell sequencing data". In: *PLoS computational biology* 11.6 (2015), e1004333.

[9] Aaron T L. Lun, Karsten Bach, and John C Marioni. "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts". In: *Genome biology* 17 (2016), pp. 1–14.

[10] Romain Lopez et al. "Deep generative modeling for single-cell transcriptomics". In: *Nature methods* 15.12 (2018), pp. 1053–1058.

[11] Christoph Hafemeister and Rahul Satija. "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression". In: *Genome biology* 20.1 (2019), p. 296.

[12] Jérémie Breda, Mihaela Zavolan, and Erik van Nimwegen. “Bayesian inference of gene expression states from single-cell RNA-seq data”. In: *Nature Biotechnology* 39.8 (Aug. 2021). Publisher: Nature Publishing Group, pp. 1008–1016.

[13] Arjun Raj et al. “Stochastic mRNA Synthesis in Mammalian Cells”. In: *PLOS Biology* 4.10 (Sept. 2006). Publisher: Public Library of Science, e309.

[14] David M. Suter et al. “Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics”. In: *Science* 332.6028 (Apr. 2011). Publisher: American Association for the Advancement of Science, pp. 472–474.

[15] Joseph S. Takahashi. “Transcriptional architecture of the mammalian circadian clock”. In: *Nature reviews. Genetics* 18.3 (Mar. 2017), pp. 164–179.

[16] Jacob J Hughey, Trevor Hastie, and Atul J Butte. “ZeitZeiger: supervised learning for high-dimensional data from an oscillatory system”. In: *Nucleic acids research* 44.8 (2016), e80–e80.

[17] Rosemary Braun et al. “Universal method for robust detection of circadian state from gene expression”. In: *Proceedings of the National Academy of Sciences* 115.39 (2018), E9247–E9256.

[18] Gang Wu et al. “A population-based gene expression signature of molecular clock phase from a single epidermal sample”. In: *Genome medicine* 12 (2020), pp. 1–12.

[19] Jacob J Hughey and Atul J Butte. “Differential phasing between circadian clocks in the brain and peripheral organs in humans”. In: *Journal of biological rhythms* 31.6 (2016), pp. 588–597.

[20] Lorenzo Talamanca, Cédric Gobet, and Felix Naef. “Sex-dimorphic and age-dependent organization of 24-hour gene expression rhythms in humans”. In: *Science* 379.6631 (2023), pp. 478–483.

[21] Denise Vlachou et al. “TimeTeller: a tool to probe the circadian clock as a multigene dynamical system”. In: *PLOS Computational Biology* 20.2 (2024), e1011779.

[22] Benjamin J Auerbach, Garret A FitzGerald, and Mingyao Li. “Tempo: an unsupervised Bayesian algorithm for circadian phase inference in single-cell transcriptomics”. In: *Nature Communications* 13.1 (2022), p. 6580.

[23] Nicole Wittenbrink et al. “High-accuracy determination of internal circadian time from a single blood sample”. In: *The Journal of clinical investigation* 128.9 (2018), pp. 3826–3839.

[24] Ron C Anafi et al. “CYCLOPS reveals human transcriptional rhythms in health and disease”. In: *Proceedings of the National Academy of Sciences* 114.20 (2017), pp. 5312–5317.

[25] Jacob J Hughey. “Machine learning identifies a compact gene set for monitoring the circadian clock in human blood”. In: *Genome medicine* 9 (2017), pp. 1–11.

[26] Takeya Kasukawa et al. “Human blood metabolite timetable indicates internal body time”. In: *Proceedings of the National Academy of Sciences* 109.37 (2012), pp. 15036–15041.

[27] Hiroki R Ueda et al. “Molecular-timetable methods for detection of body time and rhythm disorders from single-time-point genome-wide expression profiles”. In: *Proceedings of the National Academy of Sciences* 101.31 (2004), pp. 11227–11232.

[28] Bharath Ananthasubramaniam and Ramji Venkataramanan. “Time series-free rhythm profiling using COFE reveals multi-omic circadian rhythms in in-vivo human cancers”. In: *bioRxiv* (2024), pp. 2024–03.

[29] Nicholas E Phillips et al. “The circadian oscillator analysed at the single-transcript level”. In: *Molecular Systems Biology* 17.3 (Mar. 2021). Publisher: John Wiley & Sons, Ltd, e10135.

[30] Colas Droin et al. “Space-time logic of liver gene expression at sub-lobular scale”. In: *Nature Metabolism* 3.1 (Jan. 2021). Number: 1 Publisher: Nature Publishing Group, pp. 43–58.

[31] Georgi K. Marinov et al. “From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing”. In: *Genome Research* 24.3 (Mar. 2014), pp. 496–510.

[32] Björn Schwanhäusser et al. “Global quantification of mammalian gene expression control”. In: *Nature* 473.7347 (May 2011). Publisher: Nature Publishing Group, pp. 337–342.

[33] Christoph Scheiermann, Yuya Kunisaki, and Paul S. Frenette. “Circadian control of the immune system”. In: *Nature Reviews Immunology* 13.3 (Mar. 2013). Publisher: Nature Publishing Group, pp. 190–198.

[34] Hans Reinke and Gad Asher. “Crosstalk between metabolism and circadian clocks”. In: *Nature Reviews Molecular Cell Biology* 20.4 (Apr. 2019). Publisher: Nature Publishing Group, pp. 227–241.

[35] Shaon Chakrabarti and Franziska Michor. “Circadian clock effects on cellular proliferation: Insights from theory and experiments”. In: *Current Opinion in Cell Biology.* Differentiation and disease 67 (Dec. 2020), pp. 17–26.

[36] Lynne Peeples. “Medicine’s secret ingredient — it’s in the timing”. In: *Nature* 556.7701 (Apr. 2018). Bandiera_abtest: a Cg_type: News Feature Publisher: Nature Publishing Group Subject_term: Cancer, Health care, Medical research, Therapeutics, pp. 290–292.

[37] Yool Lee, Jeffrey M. Field, and Amita Sehgal. "Circadian Rhythms, Disease and Chronotherapy". In: *Journal of biological rhythms* 36.6 (Dec. 2021), pp. 503–531.

[38] Abdoulaye Karaboué et al. "Why does circadian timing of administration matter for immune checkpoint inhibitors' efficacy?" In: *British Journal of Cancer* (June 2024). Publisher: Nature Publishing Group, pp. 1–14.

[39] Ray Zhang et al. "A circadian gene expression atlas in mammals: implications for biology and medicine". In: *Proceedings of the National Academy of Sciences* 111.45 (2014), pp. 16219–16224.

[40] Marc D Ruben et al. "A database of tissue-specific rhythmically expressed human genes has potential applications in circadian medicine". In: *Science translational medicine* 10.458 (2018), eaat8806.

[41] Jarrod Shilts, Guanhua Chen, and Jacob J Hughey. "Evidence for widespread dysregulation of circadian clock progression in human cancer". In: *PeerJ* 6 (2018), e4327.

[42] Lorenzo Talamanca and Felix Naef. "How to tell time: advances in decoding circadian phase from omics snapshots". In: *F1000Research* 9 (Sept. 2020), F1000 Faculty Rev–1150.

[43] David K. Welsh et al. "Bioluminescence Imaging of Individual Fibroblasts Reveals Persistent, Independently Phased Circadian Rhythms of Clock Gene Expression". In: *Current Biology* 14.24 (Dec. 2004). Publisher: Elsevier, pp. 2289–2295.

[44] Tanya L. Leise et al. "Persistent Cell-Autonomous Circadian Oscillations in Fibroblasts Revealed by Six-Week Single-Cell Imaging of PER2::LUC Bioluminescence". In: *PLOS ONE* 7.3 (Mar. 2012). Publisher: Public Library of Science, e33334.

[45] Seung-Hee Yoo et al. "PERIOD2::LUCIFERASE real-time reporting of circadian dynamics reveals persistent circadian oscillations in mouse peripheral tissues". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.15 (Apr. 2004), pp. 5339–5346.

[46] Christian H. Gabriel et al. "Live-cell imaging of circadian clock protein dynamics in CRISPR-generated knock-in cells". In: *Nature Communications* 12.1 (June 2021). Number: 1 Publisher: Nature Publishing Group, p. 3796.

[47] Junyan Duan et al. “tauFisher predicts circadian time from a single sample of bulk and single-cell pseudobulk transcriptomic data”. In: *Nature Communications* 15.1 (May 2024). Publisher: Nature Publishing Group, p. 3840.

[48] Shaoheng Liang et al. “Latent periodic process inference from single-cell RNA-seq data”. In: *Nature Communications* 11.1 (Mar. 2020). Publisher: Nature Publishing Group, p. 1441.

[49] Shao’ang Wen et al. “Spatiotemporal single-cell analysis of gene expression in the mouse suprachiasmatic nucleus”. In: *Nature Neuroscience* 23.3 (Mar. 2020). Publisher: Nature Publishing Group, pp. 456–467.

[50] Jocelyn Y Kishi et al. “SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues”. In: *Nature methods* 16.6 (2019), pp. 533–544.

[51] Sinem K. Saka et al. “Immuno-SABER enables highly multiplexed and amplified protein imaging in tissues”. In: *Nature Biotechnology* 37.9 (Sept. 2019). Number: 9 Publisher: Nature Publishing Group, pp. 1080–1090.

[52] Joel B Sheffield. “ImageJ, a useful tool for biological image processing and analysis”. In: *Microscopy and Microanalysis* 13.S02 (2007), pp. 200–201.

[53] Carsen Stringer et al. “Cellpose: a generalist algorithm for cellular segmentation”. In: *Nature methods* 18.1 (2021), pp. 100–106.

[54] Ella Bahry et al. “RS-FISH: precise, interactive, fast, and scalable FISH spot detection”. In: *Nature Methods* 19.12 (2022), pp. 1563–1567.

[55] Shabnam Sahay et al. “An improved rhythmicity analysis method using Gaussian Processes detects cell-density dependent circadian oscillations in stem cells”. In: *Bioinformatics* 39.10 (2023), btad602.

[56] Gregor Mönke et al. “Optimal Analysis for Rhythmic Time-Dependent Biological Data-pyBOAT”. In: *Available at SSRN 3748561* (2020).

[57] Clive G. Bowsher and Peter S. Swain. “Identifying sources of variation and the flow of information in biochemical networks”. In: *Proceedings of the National Academy of Sciences* 109.20 (May 2012). Publisher: Proceedings of the National Academy of Sciences, E1320–E1328.

[58] Orly Alter, Patrick O Brown, and David Botstein. “Singular value decomposition for genome-wide expression data processing and modeling”. In: *Proceedings of the National Academy of Sciences* 97.18 (2000), pp. 10101–10106.

[59] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[60] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of machine learning research* 9.11 (2008).

[61] Yael Baran et al. "MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions". In: *Genome biology* 20 (2019), pp. 1–19.

[62] Sitara Persad et al. "SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data". In: *Nature Biotechnology* 41.12 (2023), pp. 1746–1757.

[63] Grace X. Y. Zheng et al. "Massively parallel digital transcriptional profiling of single cells". In: *Nature Communications* 8.1 (Jan. 2017). Publisher: Nature Publishing Group, p. 14049.

[64] Anna-Marie Finger et al. "Intercellular coupling between peripheral circadian oscillators by TGF-β signaling". In: *Science advances* 7.30 (2021), eabg5174.

[65] Gal Manella et al. "Circa-SCOPE: high-throughput live single-cell imaging method for analysis of circadian clock resetting". In: *Nature Communications* 12.1 (Oct. 2021). Publisher: Nature Publishing Group, p. 5903.

[66] Jordan W. Squair et al. "Confronting false discoveries in single-cell differential expression". In: *Nature Communications* 12.1 (Sept. 2021). Publisher: Nature Publishing Group, p. 5692.

[67] Monika S. Kowalczyk et al. "Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells". In: *Genome Research* 25.12 (Dec. 2015), pp. 1860–1872.

[68] Shijie C. Zheng et al. "Universal prediction of cell-cycle position using transfer learning". In: *Genome Biology* 23.1 (Jan. 2022), p. 41.

[69] Jeffrey M. Pullin and Davis J. McCarthy. "A comparison of marker gene selection methods for single-cell RNA sequencing data". In: *Genome Biology* 25.1 (Feb. 2024), p. 56.

[70] Ionut Sebastian Mihai, Sarang Chafle, and Johan Henriksson. "Representing and extracting knowledge from single-cell data". In: *Biophysical Reviews* 16.1 (Feb. 2024), pp. 29–56.

[71] Rahul Satija et al. "Spatial reconstruction of single-cell gene expression data". In: *Nature Biotechnology* 33.5 (May 2015). Publisher: Nature Publishing Group, pp. 495–502.

[72] Keren Bahar Halpern et al. "Bursty Gene Expression in the Intact Mammalian Liver". In: *Molecular Cell* 58.1 (Apr. 2015). Publisher: Elsevier, pp. 147–156.