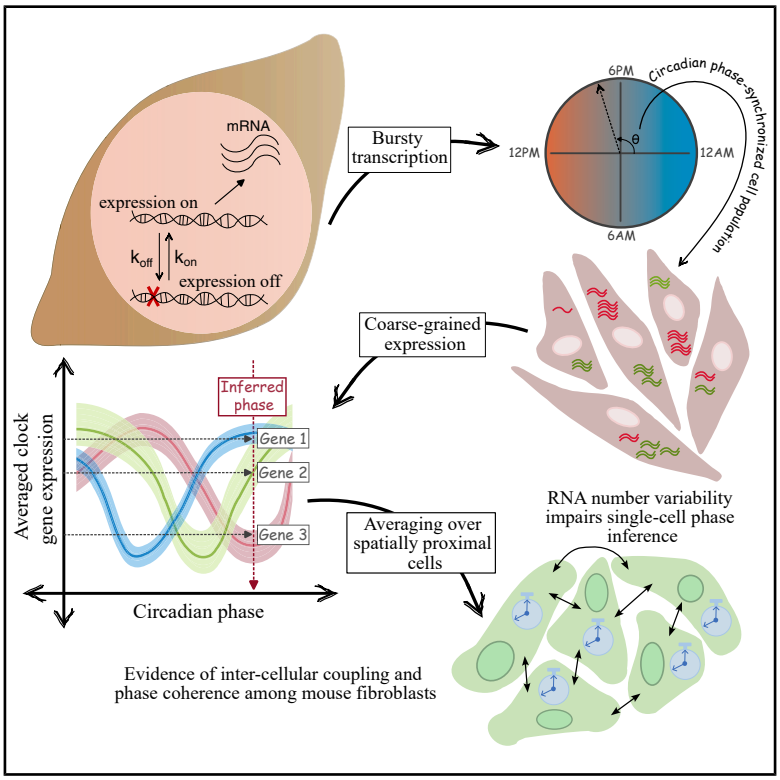
iScience

Article

Transcriptional noise sets fundamental limits to decoding circadian clock phase from single-cell RNA snapshots

Graphical abstract



Highlights

- Circadian phase inference from single-cell RNA is limited by transcriptional noise
- Only 3 core-clock genes averaged over ~70 cells give PRECISE phase inference
- Non-core-clock oscillatory genes do not improve phase inference in scRNA-seq data
- Evidence for local inter-cellular coupling among asynchronized fibroblast clocks

Authors

Anjoom Nikhat, Taniya Mandal, Nivedha Veerasubramanian, Shaon Chakrabarti

Correspondence

shaon@ncbs.res.in

In brief

Noise control; integrative aspects of cell biology; classification of bioinformatical subject; transcriptomic

Nikhat et al., 2026, iScience *29*, 115394
April 17, 2026
https://doi.org/10.1016/j.isci.2026.115394

CellPress

iScience

CellPress
OPEN ACCESS

Article

Transcriptional noise sets fundamental limits to decoding circadian clock phase from single-cell RNA snapshots

Anjoom Nikhat,[1] Taniya Mandal,[1] Nivedha Veerasubramanian,[1] and Shaon Chakrabarti[1,2,*]
[1]Simons Centre for the Study of Living Machines, National Centre for Biological Sciences, Bangalore, India
[2]Lead contact
*Correspondence: shaon@ncbs.res.in
https://doi.org/10.1016/j.isci.2026.115394

SUMMARY

The circadian clock drives rhythmic physiological variations, making accurate clock-phase measurement essential. Single-sample phase inference from bulk RNA is an exciting alternative to time-series methods, but its applicability to single cells remains unclear. Using multiplexed smFISH to quantify up to six core-clock genes in mouse fibroblasts, combined with a Gaussian process-based algorithm, we show that even with minimal technical dropouts, transcriptional noise in core-clock genes prevents reliable single-cell phase estimation. Simulations predict that measuring ~50 low-noise, clock-like oscillatory genes is required for accurate single-cell inference. In scRNA-seq data, non-core genes were too noisy, and adding hundreds of them did not improve accuracy. Remarkably, averaging smFISH counts from only three core-clock genes across ~70 cells enabled robust phase estimation. In desynchronized fibroblasts, spatially heterogeneous phases indicating local inter-cellular coupling were revealed only upon averaging, not at single-cell resolution. Coarse graining is, therefore, essential for circadian phase inference from RNA measurements alone.

INTRODUCTION

Internal clocks present across organisms have evolved in the presence of a rhythmic external environment to allow anticipation of external time. At least one of these time keepers, the circadian clock, is endogenous and operative at the level of single cells. At the molecular level, a set of transcriptional and translational feedback loops (TTFLs) combine to generate approximately 24 h oscillations of a set of core genes and proteins.[1] These molecular clocks in the suprachiasmatic nucleus (SCN) of the mammalian brain get entrained to the external light-dark cycles, resulting in an accurate internal timer with a 24-h time period. The SCN, in turn, largely controls entrainment of the millions of cells constituting various peripheral tissues, thereby providing a notion of internal time across the body.[2] Inferring the internal time (alternatively referred to as "circadian phase") is of immense importance in chronobiology, as many critical physiological processes are timed by the circadian clock,[3–5] and, hence, could be utilized for chronotherapy.[6–12] Most methods rely on measuring a proxy of the molecular clock, the chronotype, either by using qualitative questionnaires on an individual's sleep and meal patterns[13] or, more quantitatively, by biochemically detecting the onset of melatonin.[14] Recent approaches have used regression techniques based on time-series data generated by wearable devices to predict melatonin onset and internal clock phase.[15] However, issues of accuracy, logistical challenges, and masking by environmental factors have spurred recent interest in developing alternate and easier approaches to directly measuring the molecular clock phase. With rapid advances in high-throughput gene expression assays and concomitant development in machine learning algorithms, single-sample inference of circadian phase has emerged as an exciting area of research.[16] State-of-the-art methods typically use either supervised or unsupervised learning to predict the circadian phase of easily accessible tissues (like blood or skin), using RNA measurements of $\sim 10 - 15$ core-clock genes averaged over millions of cells.[17–28]

Because single-sample phase-inference techniques till date have almost solely focused on utilizing bulk RNA technologies, very little is currently understood about the possibility of phase inference from single-cell RNA snapshots. If possible, this would open up an exciting frontier of circadian phase inference, allowing for the discovery of spatial heterogeneity in endogenous clock phase within single-tissue samples, determination of the underlying cause of circadian dysfunction and dampened rhythms in diseases like cancers, and, possibility, investigation of the mechanisms of inter-cellular coupling leading to phase (de)coherence in peripheral tissues. Historically, single-cell circadian clock measurements have been performed using promoter-driven luciferase reporters in mouse tissue explants[29,30] or knock-in reporters of proteins.[29,31,32] 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 clock RNAs. While scRNA-seq has recently been used to develop phase-inference methods like Tempo[33] and tauFisher,[34] associated technical difficulties

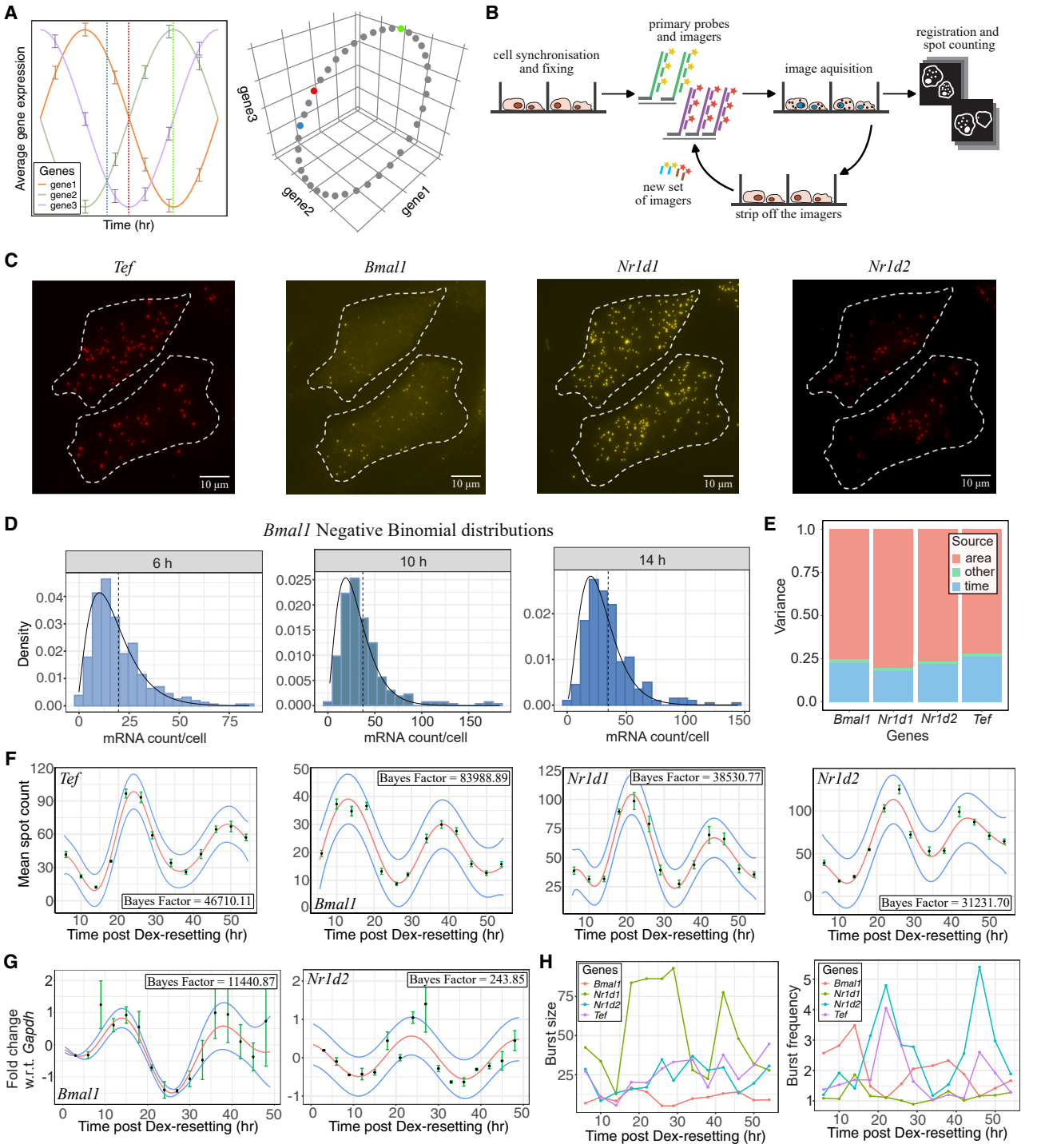


Figure 1. Single-cell gene expression recapitulates bulk properties and defines phases of the circadian clock

(A) Schematic showing how the relative levels of average gene expression uniquely define circadian phases. 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 RNA spots for the 4 clock genes, shown in the same two cells. Scale bars: 10μm.

(D) Negative binomial fits to *Bmal1* RNA 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.

(legend continued on next page)

necessitate pseudo-bulk analyses for accurate circadian phase inference, typically requiring averaging over hundreds of cells.[33–35] Although Tempo and tauFisher improve over older algorithms like Cyclops[24] and Cyclum,[36] they still give large errors in phase inference at the single-cell level. Whether this arises due to intrinsic gene-expression variability of the clock genes[37] or from technical noise invariably associated with scRNA-seq measurements remains an open question. Crucially, because scRNA-seq loses direct spatial information, it limits our ability to study phase variation across space. Some studies have utilized smFISH to obtain more accurate measurements of endogenous RNA in mouse SCN and liver, but they have been limited to a couple of clock genes and do not address the question of possibility of circadian phase inference from single cells.[38,39]

In this study, we systematically probe the limits of single-cell circadian phase decoding from RNA measurements of core-clock genes. To circumvent the major technical limitations of single-cell sequencing approaches, we utilize a recently developed multiplexed smFISH technique, SABER-FISH (signal amplification by exchange reaction-fluorescence *in situ* hybridization),[40] to obtain absolute RNA counts of up to 6 core-clock genes simultaneously in hundreds of mouse embryonic and lung fibroblasts. To optimally analyze the single-cell data, we use Gaussian processes (GPs) to develop a supervised phase-inference algorithm, PRECISE (predicting circadian phase from stochastic gene expression). Using PRECISE as well as a variety of other unsupervised algorithms, we demonstrate that biological, not technical, noise in these measured genes prevents the inference of circadian phase from single cells. Systematically adding more genes in computer simulations, using noise levels similar to the least noisy core-clock gene *Nr1d2*, allows for accurate single-cell phase inference only above 50 genes. Given that there are only about 10–20 genes of the circadian clock network that are consistently found to oscillate across tissues, our results suggest that it might be impossible to decode single-cell clock phase just from core-clock genes. Hundreds of added non-core-clock genes in a scRNA-seq dataset also did not improve phase inference, possibly due to the large technical noise associated with scRNA-seq. Remarkably, however, averaging over as few as $\sim$ 70 cells allows for accurate phase inference with $\sim$1 h error, using just 3 core-clock genes. Finally, we demonstrate how a strategy of coarse graining over a small group of spatially proximal cells can be used to resolve heterogeneous circadian phases within a single asynchronized population, providing evidence of paracrine signaling and local phase coherence among fibroblasts. These results have potential implications beyond the circadian clock cell states defined by "marker-genes" measured at the population versus single-cell level, which need not be consistent with each other and becoming comparable only after coarse graining over a small group of cells.

RESULTS

A quantitative definition of circadian clock phase from population-level RNA measurements

The oscillatory nature of core-clock gene transcripts of a synchronized population of cells, combined with the well-characterized network of genes forming the circadian clock, allows for a quantitative definition of circadian clock phase. Cells at distinct times post synchronization represent distinct circadian phases (Figure 1A, left). The circadian phase can be uniquely defined by the relative (average) RNA levels of a group of core-clock genes, which oscillate with fixed relationships between each other (Figure 1A, left).[1] It is important to note that the time post synchronization (typically referred to as zeitgeber time [ZT]) is not necessarily the same as the circadian phase, although these two quantities have often been used interchangeably in the literature. For example, the circadian clock phase velocity often varies over time, leading to unequal changes in clock phase for equally spaced ZTs. In such cases, it is important to distinguish between ZT and circadian phase, as we do in this study. Any of the oscillatory core-clock genes can be used to define the zero-phase angle. If not mentioned otherwise, throughout this work, we use the peak of the *Nr1d1* gene to define circadian phase 0 for NIH3T3 cells and the peak of the *Tef* gene to define circadian phase 0 for mouse lung fibroblast (MLG) cells. Circadian phases can alternatively be described by their angular position along the high-dimensional ellipse formed in the gene expression space (Figure 1A, right). Typically, the former formulation is used in supervised algorithms, while the latter is used in unsupervised approaches to phase inference.

A crucial aspect of this definition of circadian phase based on the core-clock RNA levels is the averaging over synchronized cellular populations. Previous algorithms have used this definition to infer circadian phase from bulk gene expression datasets.[17–28] What remains unanswered is whether the same relative RNA levels, as defined by the population averages, can be used to ascribe correct circadian phases to single cells. Answering this question is the central goal of this work.

Single-cell endogenous RNA imaging recapitulates bulk oscillations and defines phases of the circadian clock

To generate cells in these different circadian phases, we used dexamethasone (Dex) resetting to synchronize NIH3T3 cells and then captured the cells at a few discrete time points post Dex for multiplexed RNA measurements. To reduce technical noise in the measurements as far as possible, we used the recently developed SABER-FISH method to measure single-cell endogenous RNA counts.[40] This method uses long repetitive sequences called "concatemers" at the end of probes complementary to the RNA sequences of interest, which then act as platforms to hybridize many fluorescently labeled "imagers," thus amplifying the signal from each RNA molecule.[40,42] Besides

(F) Mean RNA counts and errors in the mean as functions of time. Red lines, ODeGP[41] posterior mean; blue lines, posterior variance. Bayes factors larger than 14 imply statistically significant oscillations. Data are represented as the mean $+/-$ SD.

(G) Bulk qPCR on millions of cells showing the same oscillatory patterns and phase differences for *Bmal1* and *Nr1d2* as observed in (F). Data are represented as the mean $+/-$ SD.

(H) Both burst size and burst frequency show oscillations over time (quantification of oscillations is shown in SI Section S3).

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*[40] in two cell lines used in this study— NIH3T3 and MLG. We consistently achieved ~95% co-localization in both cell types upon using two imagers simultaneously to image *Cbx5* RNA spots (Figure S1, details in Section S1), thereby demonstrating the accuracy of the measurements.

To collect cells from distinct circadian phases, we followed a protocol of adding Dex at different times and fixing cells simultaneously to minimize cell density variation across different time points (Figure S2, details in Section S1). After fixation, we performed multiplexed SABER-FISH to obtain absolute RNA counts of the 4 core-clock genes, *Bmal1*, *Tef*, *Nr1d1*, and *Nr1d2*, in about 200–250 cells per time point (Figure 1C; experimental details in STAR Methods and Sections S1 and S2). We first checked that the distributions of RNA counts for each gene across single cells followed the negative binomial (NB) 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[43] (details in Section S3) and found that a significant proportion (20%–25%) of the total variance could be ascribed to time (Figure 1E), suggesting that the average levels of gene expression across time are likely to be well separated despite the high intrinsic noise.

To confirm this assumption, 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 oscillation detection using Gaussian processes (ODeGP) (Figure 1F; details in Section S3), an oscillation detection algorithm that we previously developed and demonstrated to outperform existing methods.[41] Indeed, the oscillatory patterns recapitulated 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 RNA count distributions per time point (see Section S3 for details). We found that both parameters tend to be oscillatory over time (Figure 1H). Using ODeGP, we further quantified these oscillations and observed that both burst size and burst frequency for each of the four clock genes display strong oscillatory signatures (details of the fits and the Bayes factor value can be found in Figure S4 and Section S3). Therefore, our data suggest 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.[37]

In conclusion, the results of this section confirm that SABER-FISH can be used to accurately measure intrinsic RNA copy number noise among 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, measurement sensitivity, and Dex-synchronization protocol. Most importantly, these results suggest that the average gene expression levels at different time points post Dex resetting can be used to define unique phases of the circadian clock in NIH3T3 cells.

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

Given that the averages over ~100 cells recapitulated the 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 RNA 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 visualized the violin plots describing the distributions of absolute mRNA counts as a function of time at a single-cell level (Figure 2A, left).

If the intrinsic gene expression noise had been low, the distributions at a single-cell level would be well separated across time, and a clear oscillatory profile would have been visible. Instead, the substantial overlap between the distributions indicates high underlying noise (Figures 2A and S5), which suggests that placing single cells along an oscillatory trajectory might prove challenging. Remarkably, however, when gene expression was averaged over groups as small as 10 cells per time point, the overlap among the distributions significantly reduced, and clear oscillations emerged (Figure 2A, center and right), with correct phase differences between the genes (Figure S5). We further assessed the quality of the oscillations observed by measuring the coefficient of variation (COV) as well as Bayes factor (from the R package OdeGP). Averaging over small cell groups reduced noise (lower COV) and enhanced oscillatory signatures (higher Bayes factor) compared with single cells, as shown in Figure S6 (details in Section S5).

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) simulating cell populations with shared initial conditions and evolving them under sinusoidal dynamics with added Gaussian noise, and (2) generating mRNA count data by sampling from negative binomial distributions with burst size and frequency parameters obtained from Figure 1H, and then averaging over cells from the different populations. While approach (1) is purely for demonstrative purposes, approach (2) uses parameter estimates from our smFISH datasets and, hence, is a good representation of the underlying gene expression noise among the clock genes. Both methods demonstrate that even if oscillations are hard to detect at a single-cell level due to highly overlapping distributions (Figures 2B, right and 2C, left), averaging removes this noise and produces clear oscillations (black dots in Figure 2B). The more the underlying noise level, the more the number of cells required to observe the emergent oscillations, and, hence, it is interesting to note that only ~10 cells are sufficient to generate non-overlapping

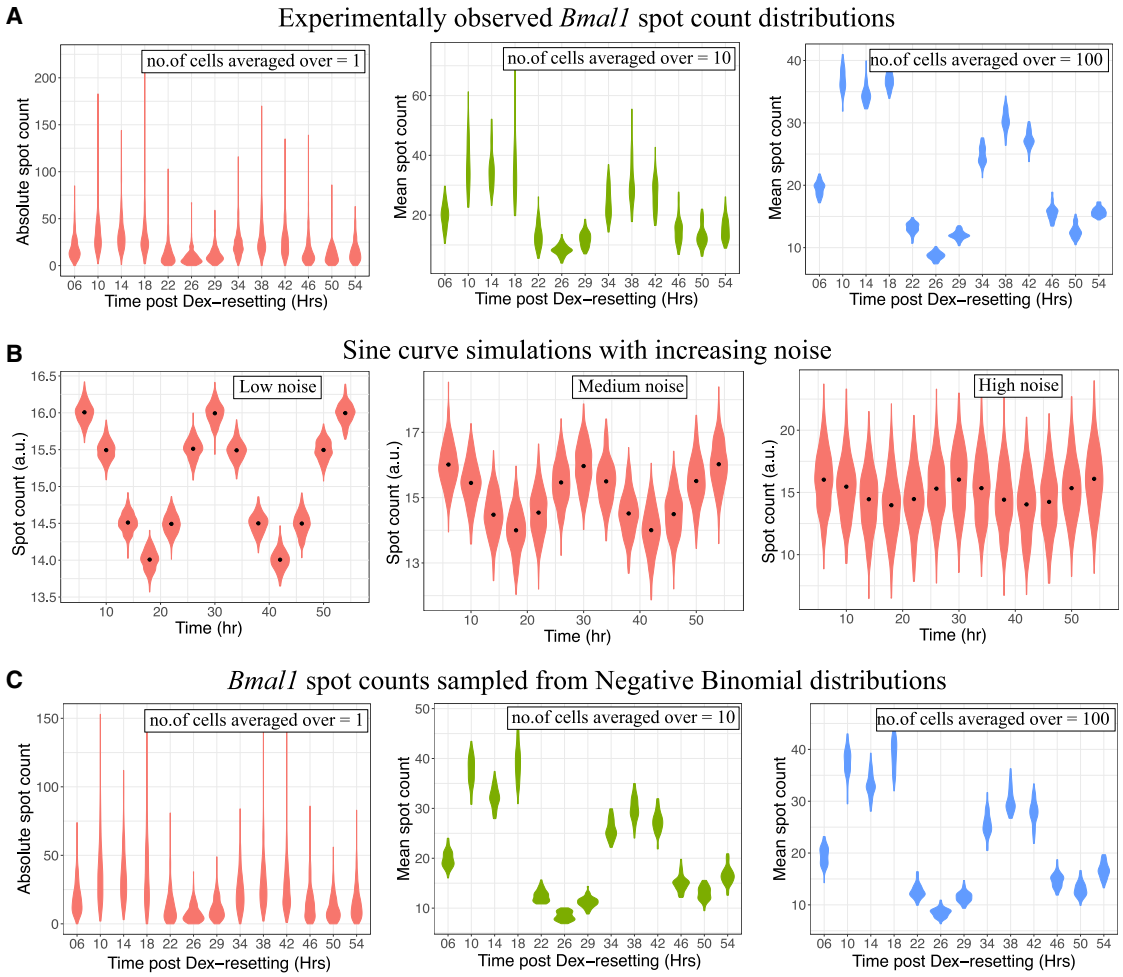


Figure 2. Emergence of oscillations and correct phase differences upon averaging over a small group of single cells

(A) Violin plots of the distributions of mean spot counts from SABER-FISH upon increasing the number of cells averaged over.

(B) Effect of noise on population-level oscillations. Populations were generated from Gaussian-sampled initial states, evolved with sinusoidal dynamics, and subjected to Gaussian noise. Violin plots show single-cell phase distributions at increasing noise levels. Despite overlap at high noise, population means (black dots) preserve the oscillatory pattern, highlighting the benefit of averaging.

(C) Violin plots of *Bmal1* spot counts generated by sampling from negative binomial distributions, using parameters obtained from fits to the SABER-FISH datasets from Figure 1D. While, at single-cell level, the distributions are overlapping, clear oscillations associated with well-separated distributions emerge upon averaging over small cellular groups, consistent with our experimental observations.

distributions characterized by clean oscillations over time in the experimental dataset (Figure 2A).

PRECISE—An accurate supervised learning algorithm for circadian phase 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 from a single RNA snapshot in terms of minimum cell and gene numbers required. Because the supervised machine learning approaches tend to outperform unsupervised ones, we first developed a supervised algorithm called PRECISE to accurately infer the circadian phase from RNA counts of a test sample. PRECISE builds on an algorithm we previously developed using GPs to learn noisy and non-stationary oscillatory trajectories.[41] Here, we provide a brief overview of the method; for more details see STAR Methods and Section S6.

The true circadian phases of the training samples are first estimated from the Dex-synchronized timeseries of normalized average spot counts (obtained by bootstrapping and normalization of the data), using a wavelet-based approach[44] (Figures 3A and S7). This process of assigning true phases allows discrimination between times post synchronization (ZTs) versus the actual circadian phase, which may be different. This is followed by learning of the optimum Bayesian posterior mean and variance of the expression levels as functions of circadian phase for each of the four genes individually. Using the posterior mean and variance, PRECISE then constructs the likelihood of a test dataset characterized by a set of clock gene expression

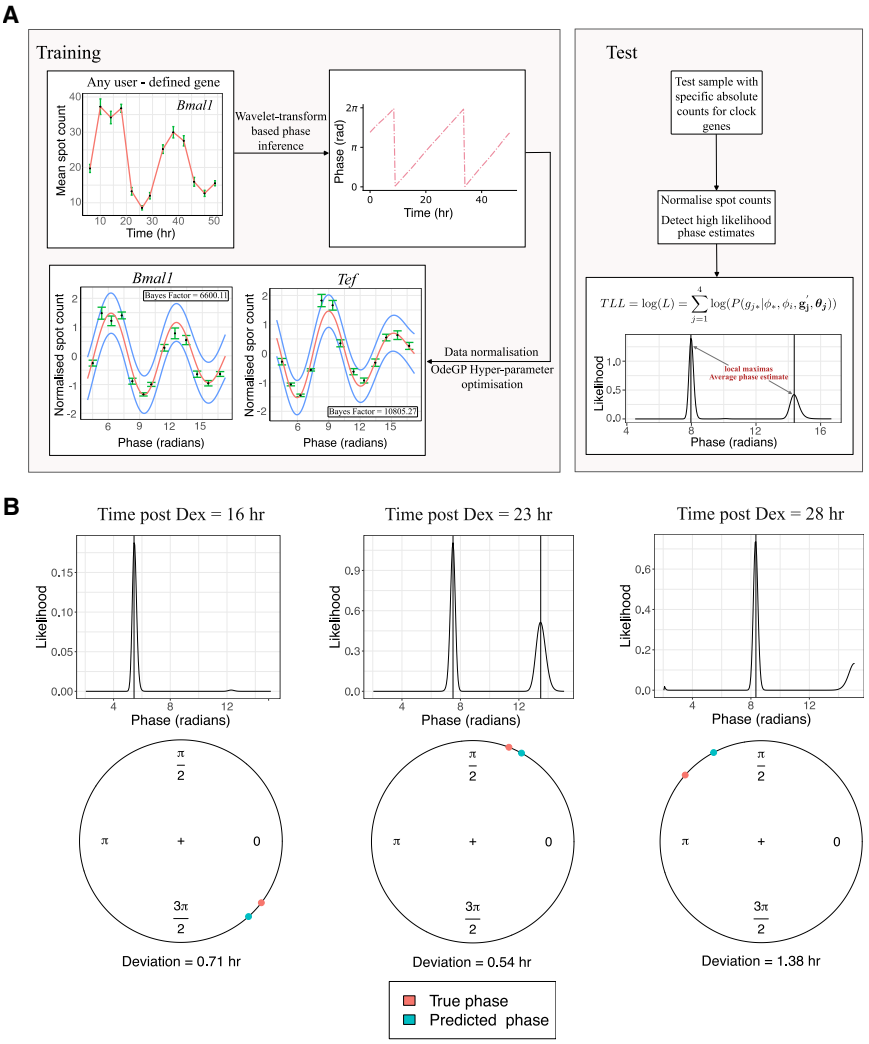


Figure 3. Accurate circadian phase inference using PRECISE

(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 normalized. PRECISE then uses Gaussian process regression to learn the optimized 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 normalizes 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 ∼ 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 phases for the 3 test cases.

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. Because PRECISE learns mapping between the gene expression and phase (as opposed to gene expression and time labels ZTs), it can handle situations where the clock phase does not progress uniformly over time.

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

Averaging core-clock RNA levels over 20–70 cells is essential for accurate circadian phase inference using PRECISE and singular value decomposition

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 h post Dex resetting—and calculated the deviation of PRECISE-inferred phase from the true phase of the cells (Δ; see STAR 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 S8). With 3 genes, however, accurate phase inference was always possible, though only after averaging over an increasing number of cells. As shown in Figure 4A (23 h test sample is shown; other time points are included in Figure S9), using only the genes *Bmal1, Nr1d1*, and *Tef*, the values of Δ almost uniformly covered the full range of 12 h when single-cell gene expression was used. This suggests that it may not be possible to decode the circadian phase with any confidence 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.[17] Δ went down further to ~ 0.5 hours upon averaging over 130 cells, and the variation in Δ shrunk to less than an hour (Figure 4A).

The empirical cumulative distribution function (eCDF) of the deviations using single-cell data was close to the diagonal line (red dotted line in Figure 4B), suggesting that these predictions are only slightly better than purely random choices of a 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 and when all 4 genes were used (Figure 4C). While all gene sets yielded Δ close to 1 h, interestingly, using 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 the true phase and the time point analyzed (Figure S10).

Like any supervised algorithm, PRECISE makes assumptions regarding the underlying model and 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 prerequisites 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.[45] Projection of samples onto the eigengene space leads to the 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 Figures 4A and 4B), 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, center) 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 only from specific time points, 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 did not cluster together as anticipated. This happens because the circadian clock's period is not exactly 24 h, and, hence, the samples do not represent the same phase. In order to quantitatively compare the performance of SVD with that of 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 than for SVD, suggesting more accurate phase estimation using our supervised learning algorithm (Figure 4E; details in Figure S11 and Section S7).

To check whether the inability to cluster single cells by phase was a result of using just 4 genes, we performed a new experiment where we measured 2 additional core-clock genes, *Cry1* and *Nifl3*, with no improvement in results (Figure S12). We then estimated the COV of the least noisy of these 6 genes (*Nr1d2* with an average COV of 0.81) and computationally simulated 10 additional clock-like genes with the same COV but with different phases (details in Section S7). This total of 16 genes represents approximately the upper limit on the number of core-clock genes that have been suggested to robustly oscillate across tissues. However, single cells remained mixed after SVD dimensionality reduction (Figure 4F), suggesting that it may not be possible to infer circadian phase from single cells by using just the core-clock genes. Interestingly, when the number of simulated genes was increased to about 50, clusters with accuracy greater than 90% emerged even at single-cell resolution (Figure 4G). This requirement of 50 genes did not change significantly when the COV was reduced by 50% to mimic the possible existence of low-noise clock-controlled genes (Figure 4G).

Finally, because previous work has demonstrated the superiority of autoencoders in cyclic phase reconstruction,[24] we also checked that an autoencoder applied to the results of SVD makes no difference to our results (details in Section S8 and Figure S13). We also demonstrated the accuracy of across-cell line phase predictions by performing inference on MLGs, when PRECISE was trained using time-course data from NIH3T3 cells (details in Section S9; Figures S14 and S15).

In summary, both the supervised and unsupervised algorithms suggest that true clock phases cannot be recovered at the single-cell level from core-clock genes alone due to the large intrinsic noise from bursty transcription. However, accurate inference with just ~ 1 hour deviation is possible with only 3 core-clock genes after averaging over a small group of cells (20–50 cells for PRECISE, ~ 70 for SVD). Interestingly, our simulations predict that about 50 oscillatory genes with noise levels similar to core-clock genes would be required for accurate phase inference from single cells.

Including non-core-clock genes is insufficient for accurate single-cell phase inference from scRNA-seq data

Because our simulations from the last section predicted that including $\sim$50 low noise, 24-h rhythmic genes would potentially allow for accurate single-cell phase inference, we next asked if this prediction could be validated using an existing scRNA-seq dataset. To this end, we analyzed a mouse skin fibroblast dataset, which provides scRNA-seq data at 6 circadian time points, $ZT2, ZT6, ZT10, ZT14, ZT18$ and $ZT22$.[34] After quality control, we retained 8,096 genes across 12,177 total cells (details in

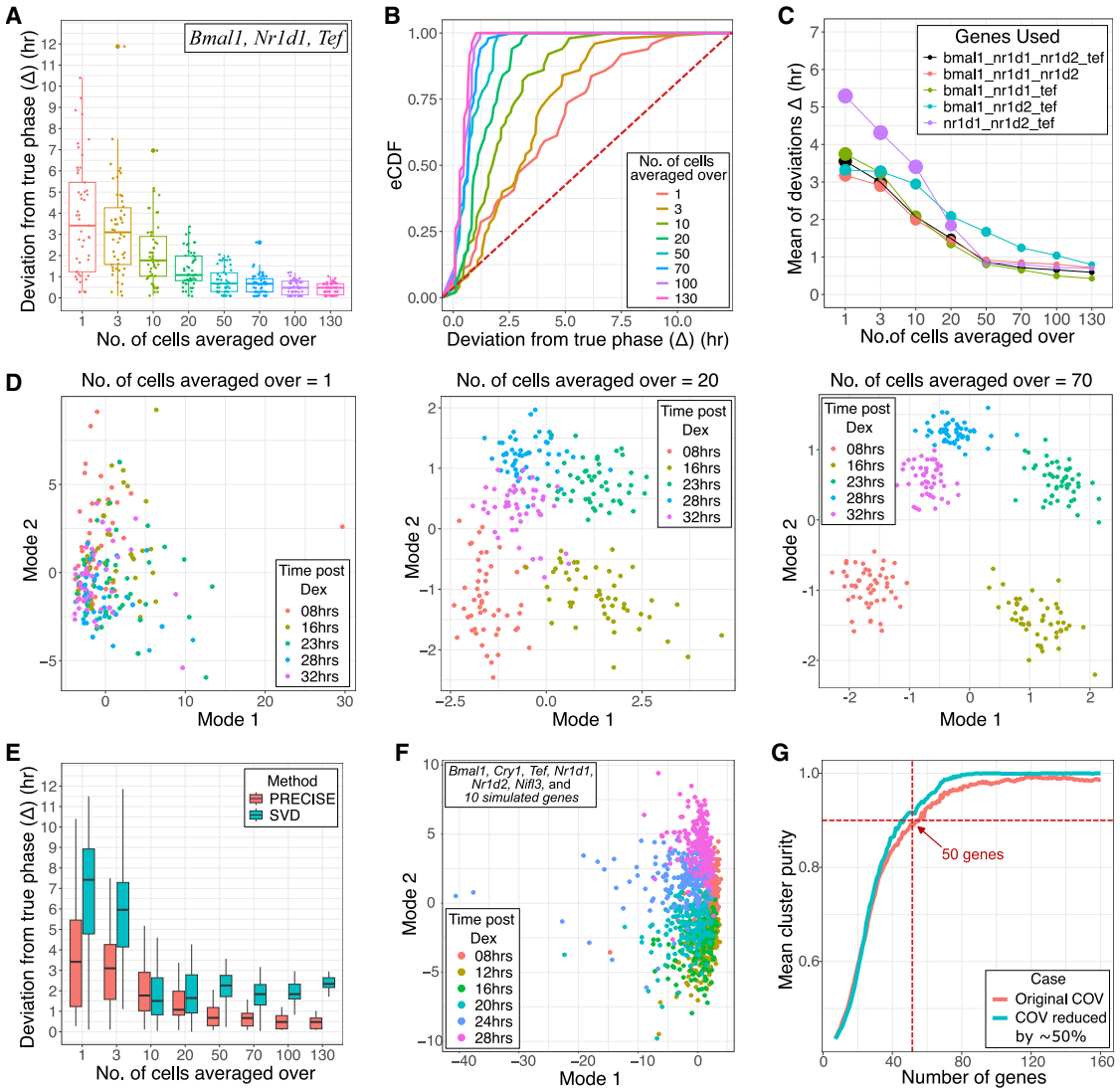


Figure 4. Coarse graining over small cellular groups is essential for accurate phase inference from core-clock RNA levels

(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 characterized by median $\Delta \sim 1$ h.

(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 gene combinations as opposed to using all 4 genes. While all gene sets yielded Δ close to 1 h, 3 genes allowed phase inference as accurately as 4 genes.

(D) Application of SVD to discover circadian phases. Left: projection of single cells in the eigengene space; cells belonging to different time points (true circadian phases) are mixed together with no elliptical structure visible. Center: 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 between the 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) SVD projection of single cells when 10 simulated genes were added to 6 measured core-clock genes.

(G) Purity of SVD clusters derived from single-cell gene expression with increasing simulated gene numbers. All simulated genes had coefficient of variations (COVs) corresponding to the least noisy measured core-clock gene, *Nr1d2* (red curve). The blue curve corresponds to simulations where genes had COV 50% less than that of *Nr1d2*.

(A)–(C) correspond to the 23-h post Dex synchronization sample of NIH3T3 cells, where true phase was assigned with respect to *Nr1d1*.

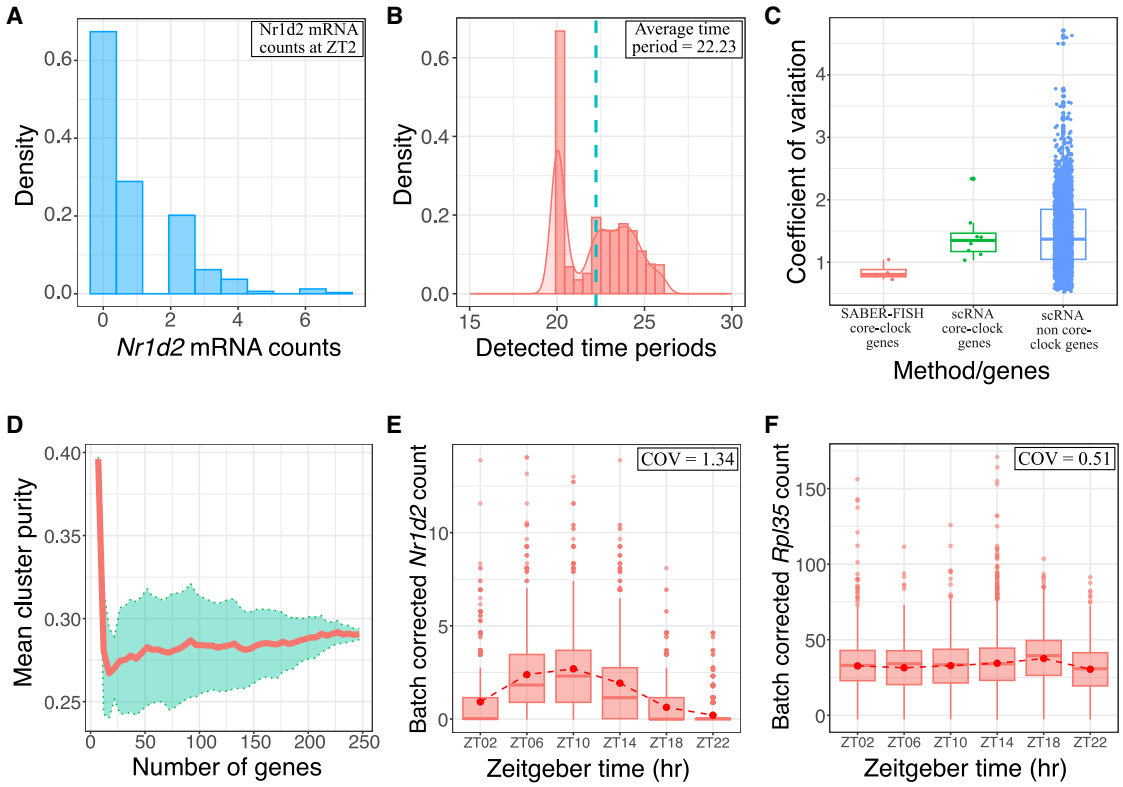


Figure 5. Including non-core-clock genes is not sufficient for accurate single-cell phase inference from scRNA-seq data

(A) Distributions of RNA counts from a scRNA-seq dataset[34] show substantial zero inflation due to technical dropouts (*Nr1d2* shown here as an example).

(B) Distribution of time periods for cyclic genes selected using MetaCycle.

(C) Coefficient of variations (COVs) calculated for core-clock genes (SABER-FISH and scRNA-seq) and non-core-clock genes (scRNA-seq). COV was calculated at every time point within the respective datasets, and the mean COV across all time points was computed.

(D) Purity of SVD clusters derived from single-cell gene expression for the increasing number of core-clock and cyclic, low-noise, non-core-clock genes.

(E) Distribution of batch corrected, single-cell *Nr1d2* (a core-clock gene) counts at the sampled time points.

(F) Distribution of batch corrected, single-cell *Rpl35* (a cyclic, low-noise, non-core-clock gene) counts at the sampled time points.

Section S10). As expected, the distribution of mRNA counts for the core-clock genes showed a substantial zero inflation due to technical dropouts (Figure 5A), which was not observed in the SABER-FISH data (Figures 1D and S3). To overcome this technical artifact in the selection of circadian genes, we averaged over all cells per time point and used MetaCycle[46] to obtain $\sim$ 2000 cyclic genes exhibiting oscillation time periods between 20 and 26 h (Figure 5B, details in Section S10). To further select the low-noise oscillatory genes, we next compared the COVs of RNA count distributions of these cyclic genes between the SABER-FISH and scRNA-seq datasets. For each gene, the COV was calculated at every time point within the respective datasets, and the mean COV across all time points was computed (Figure 5C). The COVs of core-clock genes were considerably higher in the scRNA-seq dataset than in the SABER-FISH dataset, indicating that technical noise substantially contributes to variability in gene expression measurements. The non-core-clock genes displayed a wide range of COVs, suggesting the presence of both low- and high-noise genes within this group. We selected non-core-clock genes whose COVs were lower than the SABER-FISH core-clock COVs, yielding a set of approximately 240 low-noise, circadian genes.

Using both the core-clock genes and these 240 low-noise non-core-clock genes, we then repeated the analysis of computing cluster purity following SVD-based phase inference from single-cell data. Interestingly, the inclusion of low-noise non-core-clock genes led to a decrease in cluster purity, rather than than the expected increase (Figure 5D). Note that using just the core-clock genes gave only about 40% accuracy to begin with (Figure 5D). To investigate the cause of the decrease in cluster purity, we visualized the single-cell expression distribution of a core-clock gene, *Nr1d2*, across different time points and compared it with that of the least noisy non-core-clock gene *Rpl35*. Unlike the core-clock gene that had a noisy but oscillatory profile (Figure 5E), the non-core-clock gene exhibited an almost flat profile (Figure 5F). This lack of oscillation at single-cell level was observed for other non-core-clock genes as well, which likely resulted in added noise and reduced accuracy of SVD-based phase inference. Thus, oscillations at the bulk level of non-core-clock genes do not necessarily translate into oscillations at single-cell level, and, therefore, their addition to core-clock genes can, in fact, be detrimental to phase inference.

Because single-cell phase inference via SVD failed even after including both core- and non-core-clock genes, we finally

checked the performance of our supervised learning algorithm, PRECISE. To establish a reference for evaluation, we first estimated the true circadian phase for the scRNA-seq dataset, using wavelet analysis of the mean expression profile of the core-clock gene *Tef* (see STAR Methods and Section S6). We then applied PRECISE to infer single-cell phases and quantified performance by calculating the deviation between predicted and true phases (details in Section S10). To ensure a stringent test, we evaluated PRECISE by using a curated set of 14 genes, comprising 6 core-clock genes (excluding *Cry2* as it was not classified as oscillatory by ODeGP) and 8 non-core-clock genes that were manually confirmed to exhibit circadian-like oscillations in time-course transcriptomic datasets, rather than relying solely on genes identified through computational classification. Despite increasing the number of genes, phase-inference accuracy remained low, with substantial deviations from the true phase (Figure S16). However, when expression data were averaged over small groups of cells, increasing the number of genes markedly improved PRECISE performance, indicating that averaging reduces noise and enhances the phase-inference accuracy (Figure S16).

Together, these analyses demonstrate that accurate circadian phase inference at the single-cell level depends not only on low technical noise but also on the strength of intrinsic oscillatory signals. Even if bulk profiles show rhythmicity, noise at a single-cell level can result in very weak oscillations. This appears to be true for most non-core-clock genes. Consequently, incorporating additional non-core-clock data that lack strong oscillatory behavior may introduce confounding variation, rather than enhancing phase-inference accuracy. Consistent with our conclusions from the SABER-FISH datasets, neither the unsupervised SVD approach nor the supervised PRECISE algorithm could extract accurate single-cell phases from the scRNA-seq dataset, even when hundreds of non-core-clock genes were added.

Non-linear methods fail to cluster single cells by circadian phase using core-clock genes

We next asked if, instead of the linear dimensionality reduction technique SVD, non-linear dimensionality reduction methods like UMAP (uniform manifold approximation and projection)[47] or t-SNE (t-distributed stochastic neighbor embedding)[48] would allow for better separation of single cells according to their known true circadian phases. Note that unlike PRECISE or SVD, these non-linear methods are not expected to allow for a quantitative inference of the phase because the elliptical structure (Figure 1A) will not be maintained after projection into lower dimensions. We used both UMAP and t-SNE on our absolute RNA counts to visualize the data in 2D (Figure 6A; details in STAR Methods and Section S11). Annotating the cells based on their known true circadian phases demonstrated that cells from different circadian phases were mixed together, with no clear clusters emerging, further emphasizing the fact that inherently noisy circadian gene expression limits correct circadian phase assignment at single-cell resolution. While UMAP and t-SNE are dimensionality reduction techniques, recent efforts have gone toward partitioning scRNA-seq.

In datasets to identify ''metacells'' representing distinct cell states, variability within a metacell arises from technical rather than biological sources.[49,50] We, therefore, asked if cells within a single time point or circadian phase could be discovered as a ''metacell'' and implemented a recently developed algorithm, SEACells, to infer 5 metacells in our dataset. We checked what proportion of cells from different time points were classified into a single metacell.[50] 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 6B, 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, and the 3 methods grouped very different sets of single cells into clusters, which further depended on the specific parameter choices (Figure S17, STAR Methods, and Section S11).

Given that the results in the previous section 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 UMAP and t-SNE projections in 2D clearly separated out the different time points (Figure 6C). The same was true for SEACells: each metacell identified now comprised samples from only a single time point, as shown in Figure 6D. Therefore, irrespective of the cell-state discovery algorithm we used, we could recover the true circadian states only after averaging.

Biological, not sampling, noise prevents correct single-cell phase assignments

To understand what enabled accurate phase assignment for cell groups but not single cells while using the core-clock genes, we visualized the distribution of raw spot counts for the genes across the various time points (the distributions for *Bmal1* are shown in Figure 6E; the rest are shown in Figure S18). While the means of the distributions (black dots in Figure 6E) were shifted as expected from our analysis in Figure 1F, the distributions were strongly overlapping for the different circadian phases. However, the distribution of the means was well separated as the variance of each distribution was much reduced (Figure 6F). Note that though the means of the 16 and 32 h time points were the same (Figure 6F), the presence of the other genes can allow for unique state assignments. Dimensionality reduction and cell-state discovery algorithms, with their inherent approach of clustering cells with statistically similar expression profiles, can, therefore, help discover correct phases only after averaging.

Could sampling noise have contributed to the non-identifiability of single-cell circadian phases? In sequencing-based methods like scRNA-seq, only a small fraction of the total cellular RNA (up to $\sim$ 30% with modern protocols[51]) can be captured,

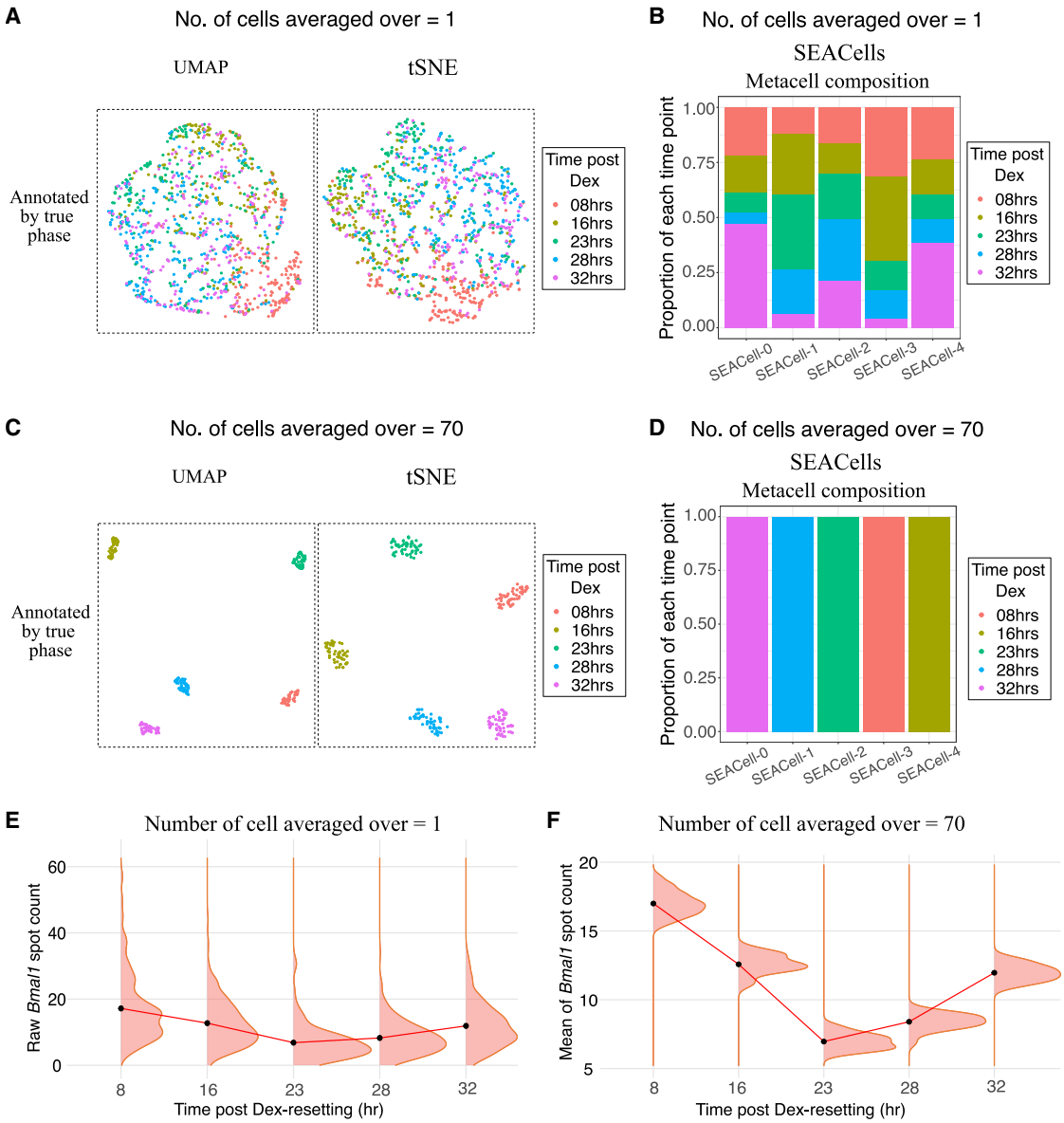


Figure 6. Transcriptional noise prevents meaningful clustering of single cells by cell-state discovery algorithms

(A) All four genes *Bmal1, Nr1d1, Nr1d2*, and *Tef* were used to obtain UMAP- and t-SNE-based dimensionality reduction. Cells annotated by the known true circadian phases (times post Dex resetting) in the UMAP and t-SNE-based projection suggest that cells from the different circadian phases were mixed together, with no clear emergent clusters.

(B) SEACells-based identification of "metacells" for our single-cell circadian clock dataset. Metacell composition shows the proportion of cells from different true circadian states that are included in each metacell. Similar to (A), each SEACell identified a metacell comprising cells from different true states.

(C and 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.

(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.

leading to sampling artifacts. 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 S12 for details). To further check if the 5% of missed transcripts could potentially lead to issues of circadian phase identification, we ran a number of simulations. We generated synthetic distributions of single-cell RNA counts for four different genes, representing two

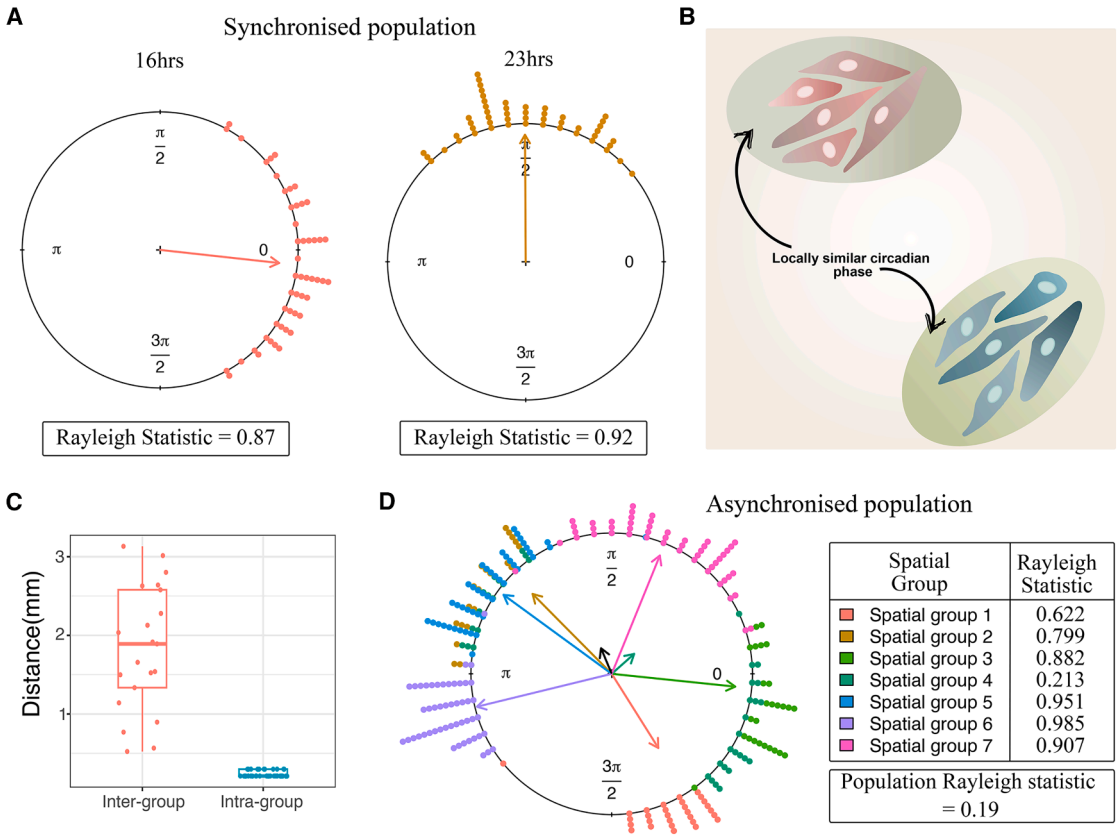


Figure 7. Averaging over locally proximal cells allows spatially resolved circadian phase inference

(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 time point; right, 23 h time point). 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 colors) in the asynchronized population. Similar to (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, a high phase coherence is observed within the different spatial groups as quantified by the high individual Rayleigh statistics (colored arrows).

distinct circadian phases. We utilized the negative binomial (NB) parameters derived from fits to our training datasets, resulting in the underlying distributions being strongly overlapping. For each of the two phases, we sampled 200 or 1,000 individual cells, each characterized 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 RNA 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 RNA per cell and annotated the samples by their true phases to analyze their clustering patterns. As shown in Figure S19, 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), phase inference was possible only after averaging.

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

Averaging RNA counts across spatially proximal fibroblasts uncovers a signature of paracrine signaling

In the last sections, we demonstrated that single-cell circadian phase inference is fundamentally limited by the transcriptional noise associated with core-clock gene expression, but averaging over small cellular groups overcomes this issue. A key question that, therefore, arises is regarding the selection of cells to be averaged over. For a synchronized population where the underlying assumption is that all cells are in the same phase, 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-synchronized test dataset, we observed that the inferred phases were localized to specific regions on the circle (Figure 7A). To quantitatively measure the phase coherence, we computed the Rayleigh statistic, whose value is close to 1 when the data are localized on a circle and 0 when the data are uniformly distributed (see details in Section S13). The high Rayleigh statistics observed for the synchronized population of cells (Figure 7A) as well as the clear separation of clusters seen earlier (Figures 6C, 6D, 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? For such asynchronous populations comprising cells from various states or phases, spatially proximal cells are likely to be representative of the same 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 7B).[52] 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.[53] We utilized these two ideas to see if spatially resolved phase inference could be performed, and if we could detect signs of paracrine signaling in mouse fibroblasts. We carried out SABER-FISH on an asynchronous population of NIH3T3 cells (details in Section S13), 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 with the intra-group distances (Figures 7C and S20). 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 visualized the inferred average phase (ϕ_{pred}) on a circle. As shown in Figures 7D and S21, both the expectations were met: (1) a uniform distribution of inferred phases was found for the entire population (black arrow corresponding to the population Rayleigh statistic), and (2) high phase coherence was observed within each spatial group, as characterized by the high individual Rayleigh statistics (colored arrows). This supports the idea that paracrine signaling between neighboring fibroblasts creates local regions of phase coherence.[52] Additionally, we inferred the phases of individual cells within the different spatial groups to examine whether cells with similar phases cluster spatially. Plotting the inferred phases relative to each cell's centroid within a spatial group revealed a heterogeneous distribution with no apparent spatial pattern (Figure S22), indicating that averaging is essential to correctly capture spatial heterogeneity in circadian phases. Note that we were limited to averaging RNA levels across just 20 spatially proximal cells in this experiment, as each field of view comprised about 70–80 cells. This is the likely cause of some amount of phase dispersion, for instance, in spatial groups 1, 2, and 4 (Figure 7D), consistent with our results from the non-spatial analysis in Figure 4D. Including more cells within a spatial neighborhood or developing more sophisticated phase-inference algorithms would reduce this dispersion and increase the accuracy of spatially resolved phase inference in the future.

DISCUSSION

Here, we undertake a systematic study of the limits of measuring circadian phase from a single snapshot of core-clock RNA levels in terms of the minimum number of cells and genes required for accurate phase inference. To circumvent the significant technical limitations of single-cell sequencing technologies that make it challenging to answer questions on fundamental limits, we utilize a highly sensitive and accurate smFISH technique called SABER-FISH[40] to measure RNA levels in single cells. Intriguingly, we find that the core-clock gene RNA ratios that are widely used to define circadian phase in population measurements cannot be directly used to ascribe meaningful circadian phases to single cells. We carefully demonstrate that this limitation arises due to biological, not technical, noise—the bursty nature of transcription results in single cells from different circadian phases having large overlap in core-clock RNA levels, even when multiple genes are measured. Remarkably, however, averaging RNA levels over small groups of 20–70 cells allows accurate phase inference with less than 1 h error, using just 3 core-clock genes. We believe that this result will provide a guideline for principled approaches to high-resolution clock-phase detection studies in the future. We also demonstrate how an approach of local averaging can be used to detect spatially heterogeneous circadian phases in an asynchronous population of cells, providing evidence for paracrine signaling among fibroblast cells.

Interestingly, computationally increasing the number of oscillatory, low-noise genes to about 50 allowed for accurate clustering by true circadian phase, even at single-cell level. Given that at maximum $\sim$, 10–20 core-clock genes are known to consistently and robustly oscillate across tissues, the addition of $\sim$30–40 robustly oscillating and low-noise, non-core-clock genes might solve the problem of single-cell phase inference. However, on analyzing a scRNA-seq dataset, we could not find non-core-clock genes that satisfied both the criteria of low-noise and robust oscillations. On the contrary, including hundreds of circadian non-core-clock genes resulted in worse single-cell phase accuracy compared with that when including just the core-clock genes.

The main prediction of this study that inferring circadian phase is not possible from single cells using just the core-clock RNA levels, although feasible using population measurements, may appear surprising at first. However, it is increasingly becoming clear that functional cell states need not emerge from RNA levels alone but likely from a combination of other cellular components like protein levels, the epigenome, and chromatin architecture.[54] With less noise associated with higher copy numbers, measurement of only a few protein levels could conceivably allow for single-cell level determination of circadian phases. However, if only RNA levels can be experimentally accessed, coarse graining may become necessary, as our work suggests. Interestingly,

we found a conceptually similar result in a different context—in an older study using scRNA-seq with spike-in controls to distinguish between biological and technical noise.[55] This study demonstrated that to recover the same transcriptome and number of measurable genes as obtained from bulk RNA-seq measurements, RNA levels from groups of 30–100 cells needed to be averaged over.[55] This number is remarkably similar to our finding of 20–70 cells, though in a very different context. These results together suggest a potentially general principle beyond circadian clock phase inference, even if technical noise can be minimized: averaging over a small group of cells as opposed to using single-cell RNA levels might be necessary for a robust definition of cell states that is consistent between the population- and single-cell levels.

Limitations of the study

Our study has a number of limitations that will need to be clarified in future work. First and foremost, our main results have been derived from noise levels in cell lines. While transcriptional bursts have been reported in tissues,[56] there might be additional noise reduction mechanisms in the *in vivo* context that may change the level of coarse graining or gene numbers predicted by our current results. Though the scRNA-seq data analysis on mouse tissue supported our results derived from cell lines, the large technical noise in scRNA-seq still remains a confounding factor. Therefore, it may still be possible, with improved techniques that reduce dropouts, to identify low-noise, non-core-clock genes that would allow single-cell resolved phase inference. Second, Dex resetting might be imperfect in synchronizing 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. Because 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.[53] If there were a number of other states mixed in due to incomplete Dex action, one would have expected one large cluster, along with some small clusters, upon phase inference. We did not see these behaviors and observed that single-cell phase inferences were uniformly distributed (Figure 4A), and clusters were well mixed in the true states (Figures 5A and 5B), thus ruling out any major effects of incomplete Dex synchronization. Finally, it remains possible that more powerful computational algorithms developed in the future could push down the limits we discovered in this study. For example, rapid developments in representation learning methods or the methods using RNA velocity could potentially be leveraged to challenge the limits identified here.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Dr. Shaon Chakrabarti (shaon@ncbs.res.in).

Materials availability

This study did not generate any new reagents.

Data and code availability

- SABER-FISH datasets generated in the paper will be available with the article. The various files for the same are labeled as follows:
  - NIH3T3 48 h training data for 4 genes → Data_S1_training_Data_NIH3T3_4genes.xlsx.
  - NIH3T3 24-h test data for 4 genes → Data_S2_test_Data_NIH3T3_4genes.xlsx.
  - MLG 24 h test data for 4 genes → Data_S3_test_Data_MLG_4genes.xlsx.
  - NIH3T3 24 h data for 6 genes → Data_S4_test_Data_NIH3T3_6genes.xlsx.
  - Asynchronous NIH3T3 data for 4 genes → Data_S5_Asynchronous_NIH3T3_4genes.xlsx.
- Publicly available dataset analyzed in the paper is listed in key resources table.
- All codes used in the manuscript are available at https://github.com/Shaonlab/Circadian-phase-estimation—PRECISE.
- Any additional information required is available from the lead contact upon request.

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) via Department of Atomic Energy, Government of India, project identification no. RTI 4006.

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.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Cell culture
- METHOD DETAILS
  - SABER-FISH for multiplexed RNA imaging
  - Image analysis to quantify single-cell gene expression
  - Bulk gene expression measurements using qPCR
  - PRECISE: PREdicting CIrcadian phase from stochastic gene expression
  - Assignment of ‘true’ circadian phase to time-points post Dex-resetting
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Calculation of deviation of inferred phase from true circadian phase
  - Unsupervised dimensionality reduction and cell-state discovery algorithms
  - scRNA-seq data analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2026.115394.

Received: June 19, 2025
Revised: October 23, 2025
Accepted: March 13, 2026
Published: March 19, 2026

REFERENCES

1. Takahashi, J.S. (2017). Transcriptional architecture of the mammalian circadian clock. Nat. Rev. Genet. *18*, 164–179.
2. Reppert, S.M., and Weaver, D.R. (2002). Coordination of circadian timing in mammals. Nature *418*, 935–941.
3. Scheiermann, C., Kunisaki, Y., and Frenette, P.S. (2013). Circadian control of the immune system. en. Nat. Rev. Immunol. *13*, 190–198.
4. Reinke, H., and Asher, G. (2019). Crosstalk between metabolism and circadian clocks. Nat. Rev. Mol. Cell Biol. *20*, 227–241.
5. Chakrabarti, S., and Michor, F. (2020). Circadian clock effects on cellular proliferation: Insights from theory and experiments. Curr. Opin. Cell Biol. *67*, 17–26.
6. Peeples, L. (2018). Medicine's secret ingredient — it's in the timing. Nature *556*, 290–292.
7. Lee, Y., Field, J.M., and Sehgal, A. (2021). Circadian Rhythms, Disease and Chronotherapy. J. Biol. Rhythms *36*, 503–531.
8. Karaboué, A., Innominato, P.F., Wreglesworth, N.I., Duchemann, B., Adam, R., and Lévi, F.A. (2024). Why does circadian timing of administration matter for immune checkpoint inhibitors' efficacy? Br. J. Cancer *131*, 783–796.
9. Zhang, R., Lahens, N.F., Ballance, H.I., Hughes, M.E., and Hogenesch, J.B. (2014). A circadian gene expression atlas in mammals: implications for biology and medicine. Proc. Natl. Acad. Sci. USA *111*, 16219–16224.
10. Ruben, M.D., Wu, G., Smith, D.F., Schmidt, R.E., Francey, L.J., Lee, Y.Y., Anafi, R.C., and Hogenesch, J.B. (2018). A database of tissue-specific rhythmically expressed human genes has potential applications in circadian medicine. Sci. Transl. Med. *10*, eaat8806.
11. Shilts, J., Chen, G., and Jacob, J.H. (2018). Evidence for widespread dysregulation of circadian clock progression in human cancer. PeerJ *6*, e4327.
12. Nikhat, A., Shaikh, A., and Chakrabarti, S. (2025). Combining lineage correlations and a small molecule inhibitor to detect circadian control of the cell cycle. iScience *28*, 112269.
13. Roenneberg, T., Pilz, L.K., Zerbini, G., and Winnebeck, E.C. (2019). Chronotype and Social Jetlag: A (Self-) Critical Review. Biology *8*, 54.
14. Murray, J.M., Stone, J.E., Abbott, S.M., Bjorvatn, B., Burgess, H.J., Cajochen, C., Dekker, J.J., Duffy, J.F., Epstein, L.J., Garbazza, C., et al. (2024). A Protocol to Determine Circadian Phase by At-Home Salivary Dim Light Melatonin Onset Assessment. eng Journal of Pineal Research. J. Pineal Res. *76*, e12994.
15. Stone, J.E., Postnova, S., Sletten, T.L., Rajaratnam, S.M., and Phillips, A.J. (2020). Computational approaches for individual circadian phase prediction in field settings. Curr. Opin. Syst. Biol. *22*, 39–51.
16. Talamanca, L., and Naef, F. (2020). How to tell time: advances in decoding circadian phase from omics snapshots. F1000Res. *9*, F1000.Faculty.Rev.1150.
17. Jacob, J.H., Hastie, T., and Butte, A.J. (2016). ZeitZeiger: supervised learning for high-dimensional data from an oscillatory system. Nucleic Acids Res. *44*, e80.
18. Braun, R., Kath, W.L., Iwanaszko, M., Kula-Eversole, E., Abbott, S.M., Reid, K.J., Zee, P.C., and Allada, R. (2018). Universal method for robust detection of circadian state from gene expression. Proc. Natl. Acad. Sci. USA *115*, E9247–E9256.
19. Wu, G., Ruben, M.D., Francey, L.J., Smith, D.F., Sherrill, J.D., Oblong, J.E., Mills, K.J., and Hogenesch, J.B. (2020). A population-based gene expression signature of molecular clock phase from a single epidermal sample. Genome Med. *12*, 73.
20. Jacob, J.H., and Butte, A.J. (2016). Differential phasing between circadian clocks in the brain and peripheral organs in humans. J. Biol. Rhythm. *31*, 588–597.
21. Talamanca, L., Gobet, C., and Naef, F. (2023). Sex-dimorphic and age-dependent organization of 24-hour gene expression rhythms in humans. Science *379*, 478–483.
22. Vlachou, D., Veretennikova, M., Usselmann, L., Vasilyev, V., Ott, S., Bjarnason, G.A., Dallmann, R., Levi, F., and Rand, D.A. (2024). TimeTeller: a tool to probe the circadian clock as a multigene dynamical system. PLoS Comput. Biol. *20*, e1011779.
23. Wittenbrink, N., Ananthasubramaniam, B., Münch, M., Koller, B., Maier, B., Weschke, C., Bes, F., de Zeeuw, J., Nowozin, C., Wahnschaffe, A., et al. (2018). High-accuracy determination of internal circadian time from a single blood sample. J. Clin. Investig. *128*, 3826–3839.
24. Anafi, R.C., Francey, L.J., Hogenesch, J.B., and Kim, J. (2017). CYCLOPS reveals human transcriptional rhythms in health and disease. Proc. Natl. Acad. Sci. USA *114*, 5312–5317.
25. Jacob, J.H. (2017). Machine learning identifies a compact gene set for monitoring the circadian clock in human blood. Genome Med. *9*, 1–11.
26. Kasukawa, T., Sugimoto, M., Hida, A., Minami, Y., Mori, M., Honma, S., Honma, K., Mishima, K., Soga, T., and Ueda, H.R. (2012). Human blood metabolite timetable indicates internal body time. Proc. Natl. Acad. Sci. USA *109*, 15036–15041.
27. Ueda, H.R., Chen, W., Minami, Y., Honma, S., Honma, K., Iino, M., and Hashimoto, S. (2004). Molecular-timetable methods for detection of body time and rhythm disorders from single-time-point genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *101*, 11227–11232.
28. Ananthasubramaniam, B., and Venkataramanan, R. (2025). Time series-free rhythm profiling using COFE reveals multi-omic circadian rhythms in in-vivo human cancers. PLoS Biol. *23*, e3003196.
29. Welsh, D.K., Yoo, S.H., Liu, A.C., Takahashi, J.S., and Kay, S.A. (2004). Bioluminescence Imaging of Individual Fibroblasts Reveals Persistent, Independently Phased Circadian Rhythms of Clock Gene Expression. English. Curr. Biol. *14*, 2289–2295.
30. Leise, T.L., Wang, C.W., Gitis, P.J., and Welsh, D.K. (2012). Persistent Cell-Autonomous Circadian Oscillations in Fibroblasts Revealed by Six-Week Single-Cell Imaging of PER2::LUC Bioluminescence. PLoS One *7*, e33334.
31. Yoo, S.-H., Yamazaki, S., Lowrey, P.L., Shimomura, K., Ko, C.H., Buhr, E.D., Siepka, S.M., Hong, H.K., Oh, W.J., Yoo, O.J., et al. (2004). PERIOD2::LUCIFERASE real-time reporting of circadian dynamics reveals persistent circadian oscillations in mouse peripheral tissues. Proc. Natl. Acad. Sci. USA *101*, 5339–5346.
32. Gabriel, C.H., Del Olmo, M., Zehtabian, A., Jäger, M., Reischl, S., van Dijk, H., Ulbricht, C., Rakhymzhan, A., Korte, T., Koller, B., et al. (2021). Live-cell imaging of circadian clock protein dynamics in CRISPR-generated knock-in cells. Nat. Commun. *12*, 3796.
33. Auerbach, B.J., FitzGerald, G.A., and Li, M. (2022). Tempo: an unsupervised Bayesian algorithm for circadian phase inference in single-cell transcriptomics. Nat. Commun. *13*, 6580.
34. Duan, J., Ngo, M.N., Karri, S.S., Tsoi, L.C., Gudjonsson, J.E., Shahbaba, B., Lowengrub, J., and Andersen, B. (2024). tauFisher predicts circadian time from a single sample of bulk and single-cell pseudobulk transcriptomic data. Nat. Commun. *15*, 3840.
35. Tong, C.Y., Li, C., Hurni, C., Jacq, A., Nie, X.Y., Guy, C.R., Suh, J.H., Wong, R.K.W., Merlin, C., Naef, F., et al. (2025). Single-Cell Multiomic

Analysis of Circadian Rhythmicity in Mouse Liver. Preprint at bioRxiv. https://doi.org/10.1101/2025.04.03.647044.

36. Shaoheng, L., Fang, W., Jincheng, H., and Ken, C. (2020). Latent periodic process inference from single-cell RNA-seq data. Nat. Commun. *11*, 1441.
37. Phillips, N.E., Hugues, A., Yeung, J., Durandau, E., Nicolas, D., and Naef, F. (2021). The circadian oscillator analysed at the single-transcript level. Mol. Syst. Biol. *17*, e10135.
38. Wen, S., Ma, D., Zhao, M., Xie, L., Wu, Q., Gou, L., Zhu, C., Fan, Y., Wang, H., and Yan, J. (2020). Spatiotemporal single-cell analysis of gene expression in the mouse suprachiasmatic nucleus. Nat. Neurosci. *23*, 456–467.
39. Droin, C., Kholtei, J.E., Bahar Halpern, K., Hurni, C., Rozenberg, M., Muvkadi, S., Itzkovitz, S., and Naef, F. (2021). Space-time logic of liver gene expression at sub-lobular scale. en. Nat. Metab. *3*, 43–58.
40. Kishi, J.Y., Lapan, S.W., Beliveau, B.J., West, E.R., Zhu, A., Sasaki, H.M., Saka, S.K., Wang, Y., Cepko, C.L., and Yin, P. (2019). SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues. Nat. Methods *16*, 533–544.
41. Sahay, S., Adhikari, S., Hormoz, S., and Chakrabarti, S. (2023). An improved rhythmicity analysis method using Gaussian Processes detects cell-density dependent circadian oscillations in stem cells. Bioinformatics *39*, btad602.
42. Saka, S.K., Wang, Y., Kishi, J.Y., Zhu, A., Zeng, Y., Xie, W., Kirli, K., Yapp, C., Cicconet, M., Beliveau, B.J., et al. (2019). Immuno-SABER enables highly multiplexed and amplified protein imaging in tissues. Nature Biotechnology *37*, 1080–1090.
43. Bowsher, C.G., and Swain, P.S. (2012). Identifying sources of variation and the flow of information in biochemical networks. Proc. Natl. Acad. Sci. USA *109*, E1320–E1328.
44. Gregor, M., Frieda, S., Christoph, S., and Adrian, E.G. (2020). Optimal Analysis for Rhythmic Time-Dependent Biological Data-pyBOAT. https://doi.org/10.2139/ssrn.3748561.
45. Alter, O., Brown, P.O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. USA *97*, 10101–10106.
46. Wu, G., Anafi, R.C., Hughes, M.E., Kornacker, K., and Hogenesch, J.B. (2016). MetaCycle: an integrated R package to evaluate periodicity in large scale data. Bioinformatics *32*, 3351–3353. https://academic.oup.com/bioinformatics/article-pdf/32/21/3351/7889685/btw405.pdf.
47. McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.03426.
48. Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*, 11.
49. Baran, Y., Bercovich, A., Sebe-Pedros, A., Lubling, Y., Giladi, A., Chomsky, E., Meir, Z., Hoichman, M., Lifshitz, A., and Tanay, A. (2019). MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol. *20*, 206–219.
50. Persad, S., Choo, Z.N., Dien, C., Sohail, N., Masilionis, I., Chaligné, R., Nawy, T., Brown, C.C., Sharma, R., Pe'er, I., et al. (2023). SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. Nat. Biotechnol. *41*, 1746–1757.
51. Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, 14049.
52. Finger, A.M., Jäschke, S., Del Olmo, M., Hurwitz, R., Granada, A.E., Herzel, H., and Kramer, A. (2021). Intercellular coupling between peripheral circadian oscillators by TGF-β signaling. Sci. Adv. *7*, eabg5174.
53. Manella, G., Aizik, D., Aviram, R., Golik, M., and Asher, G. (2021). Circa-SCOPE: high-throughput live single-cell imaging method for analysis of circadian clock resetting. Nat. Commun. *12*, 5903.
54. Rafelski, S.M., and Theriot, J.A. (2024). Establishing a conceptual framework for holistic cell states and state transitions. Cell *187*, 2633–2651.
55. Marinov, G.K., Williams, B.A., McCue, K., Schroth, G.P., Gertz, J., Myers, R.M., and Wold, B.J. (2014). From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. Genome Res. *24*, 496–510.
56. Bahar Halpern, K., Tanami, S., Landen, S., Chapal, M., Szlak, L., Hutzler, A., Nizhberg, A., and Itzkovitz, S. (2015). Bursty Gene Expression in the Intact Mammalian Liver. Mol. Cell *58*, 147–156.
57. Joel, B.S. (2007). ImageJ, a useful tool for biological image processing and analysis. Microsc. Microanal. *13*, 200–201.
58. Stringer, C., Wang, T., Michaelos, M., and Pachitariu, M. (2021). Cellpose: a generalist algorithm for cellular segmentation. Nat. Methods *18*, 100–106.
59. Bahry, E., Breimann, L., Zouinkhi, M., Epstein, L., Kolyvanov, K., Mamrak, N., King, B., Long, X., Harrington, K.I.S., Lionnet, T., et al. (2022). RS-FISH: precise, interactive, fast, and scalable FISH spot detection. Nat. Methods *19*, 1563–1567.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Analyzed dataset	Duan et al.[34]	N/A
Software and algorithms		
PRECISE	This paper	https://github.com/Shaonlab/Circadian-phase-estimation—PRECISE
Image analysis codes	This paper	https://github.com/Shaonlab
Other		
NIH3T3 4 genes training data	This paper	(Data_S1_training_Data_NIH3T3_4genes.xlsx)
NIH3T3 4 genes test data	This paper	(Data_S2_test_Data_NIH3T3_4genes.xlsx)
MLG 4 genes test data	This paper	(Data_S3_test_Data_MLG_4genes.xlsx)
NIH3T3 6 genes test data	This paper	(Data_S4_test_Data_NIH3T3_6genes.xlsx)
NIH3T3 4 genes Asynchronous data	This paper	(Data_S5_Asynchronous_NIH3T3_4genes.xlsx)

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

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. RM10432-500 ML) 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.

METHOD DETAILS

SABER-FISH for multiplexed RNA imaging

To visualize endogenous RNA of different clock genes that are produced in low numbers, we utilised a recently developed technique Signal Amplification by Exchange Reaction - Fluorescence *in-situ* Hybridisation (SABER-FISH),[40] with all reagent concentrations and reaction times taken from the updated protocols provided in,[42] To generate the training dataset we performed multiplexed imaging of 4 clock genes – *Tef, Bmal1, Nr1d1* and *Nr1d2* for fixed monolayers of cells over every 4 h for 48 h after Dexamethasone treatment. To generate NIH3T3cell line test dataset we obtained the counts of the same four clock genes over time points – 8, 16, 23, 28 and 32 h post Dex-resetting. To generate MLG cell line test datasets we measured the 4 clock genes included in training data across time points – 8,16, 23, 28 h post Dex-resetting. In a separate experiment performed on the NIH3T3cell line, we obtained the absolute RNA counts of 6 clock genes – *Tef, Bmal1, Nr1d1, Cry1, Nifl3* and *Nr1d2* over 6 different time points post Dex-resetting – 8, 12, 16, 20, 24 and 28 h. 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 60× 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.[57,58] We detected and extracted the spot coordinates for the different genes from maximum Z-projected images using the RS-FISH plugin in ImageJ.[59] 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 RNA 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 an algorithm, PRECISE (PREdicting CIrcadian phase from Stochastic gene Expression), to infer the circadian phase of a test sample from RNA measurements in a supervised manner. The input training data for PRECISE comprises the normalized mean RNA levels as functions of circular phase values (obtained from time post Dexamethasone synchronisation expression data), as well as the errors of the normalized 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,[41] 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[41]) 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 h 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 S7 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, as described in Section S6. The bootstrapped data is then subjected to normalization by which the mean and standard deviation over time was made 0 and 1 respectively as described in Section S6. This normalized input was utilised by ODeGP[41] to determine the posterior mean fit to the data, allowing extraction of the normalized 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[44] for determining phase values. pyBOAT is an algorithm that uses wavelet-transforms to determine the phase of the different samples as functions of time.[44] 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 $(i = 1, 2, \ldots, 13)$ while the phase estimates of the test samples are denoted ϕ_{true}. More details and parameter values of softwares used are provided in Figure S7 and Section S6.

QUANTIFICATION AND STATISTICAL ANALYSIS

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):

$$\delta = |\phi_{pred} - \phi_{true}|, \qquad \text{(Equation 1)}$$

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

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

Unsupervised dimensionality reduction and 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 2-node bottleneck layer (Section S8), UMAP, t-SNE and SEACells (Section S11) to discover circadian clock states. Details of each method are provided in the supplemental information as indicated.

scRNA-seq data analysis

We analyzed data from a recent paper[34] where scRNA-seq was performed on mouse skin samples at 6 different time points – ZT2, ZT6, ZT10, ZT14, ZT18 and ZT22 respectively. For each time point 3 replicates were collected. We performed QC on the dataset using the seurat package in R. We then used marker gene expression to select only the expression data from skin fibroblasts. Selection of the cyclic genes was performed using pseudo-bulk RNA counts as mentioned in the original paper.[34] Pseudo-bulk counts were calculated by summing up the RNA counts for each gene from all cells in a particular replicate. The meta2d function from the MetaCycle package in R was then applied to the pseudobulk data to identify rhythmic genes. Genes with a JTK_pvalue < 0.05 were considered significantly rhythmic. Using the more stringent q-value to correct for multiple hypothesis testing left essentially no oscillatory genes from the dataset. Further details of the scRNA-seq data analysis can be found in Section S10.