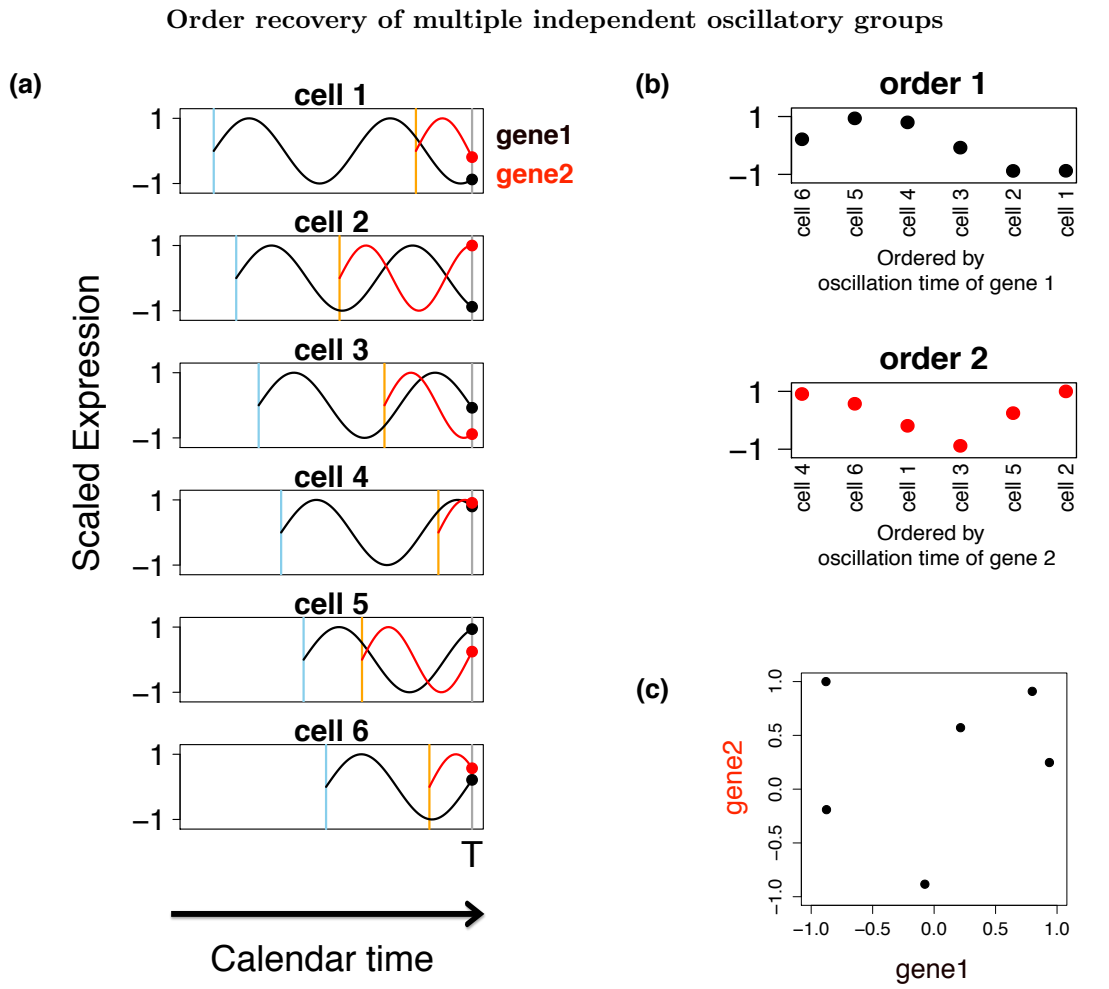
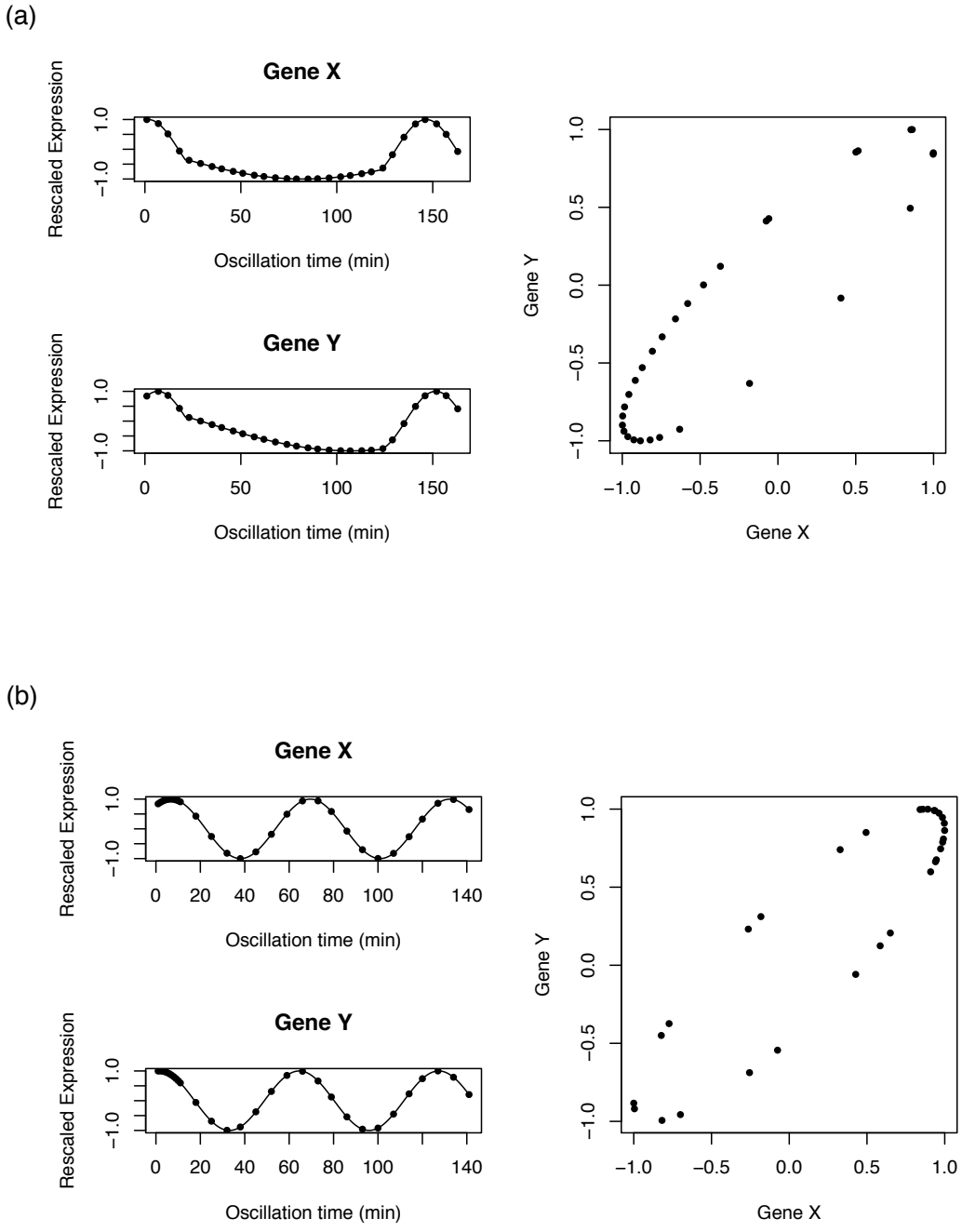
Supplementary note for "Oscope identifies oscillatory genes in unsynchronized single cell RNA-seq experiments"



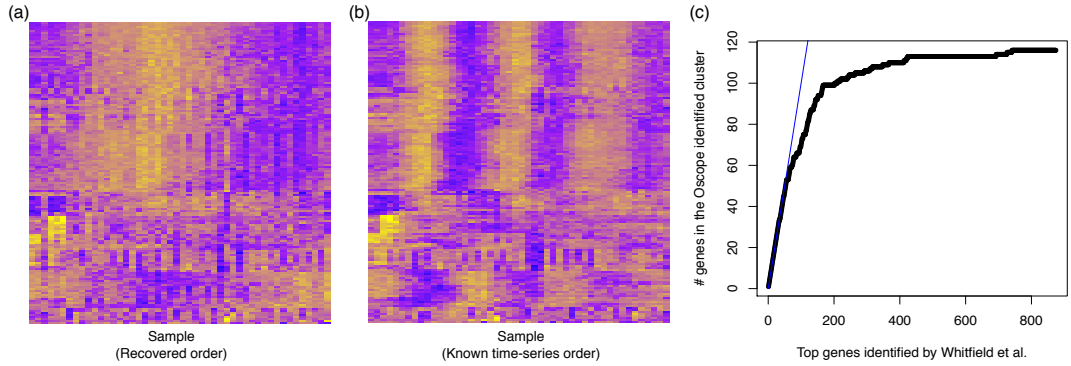
Supplementary Figure 1: Panel (a) shows two oscillatory genes, from independent oscillatory groups. The x-axis shows the calendar time. The oscillation start times of genes 1 and 2 are marked in blue and orange, respectively. Panel (b) shows the recovered base cycle using gene 1 (upper panel) and gene 2 (lower panel). Panel (c) shows the expression scatter plot of gene 1 vs. gene 2.

Oscillation with varying speed or partial synchronization



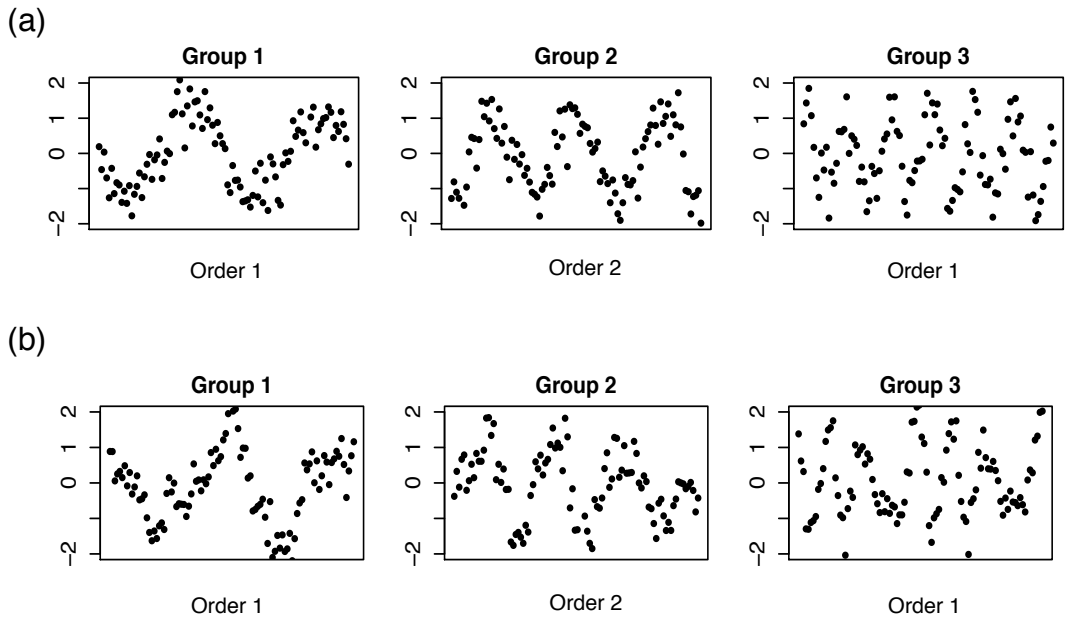
Supplementary Figure 2: The left two panels of (a) show two co-regulated oscillatory genes over oscillation time. Both genes follow a sinusoid but have a slower progression in the interval between 20min and 120min. Thirty cells from an unsynchronized population were collected at the same calendar time. Each cell's profile at the given collecting time is marked as a black dot. The right panel shows the co-regulation structure between these two genes. The left two panels of (b) show another two co-regulated oscillatory genes with constant oscillation speed. Thirty cells from a partially synchronized population were collected at the same calendar time. The right panel shows the co-regulation structure between these two genes.

Case study results on Whitfield data

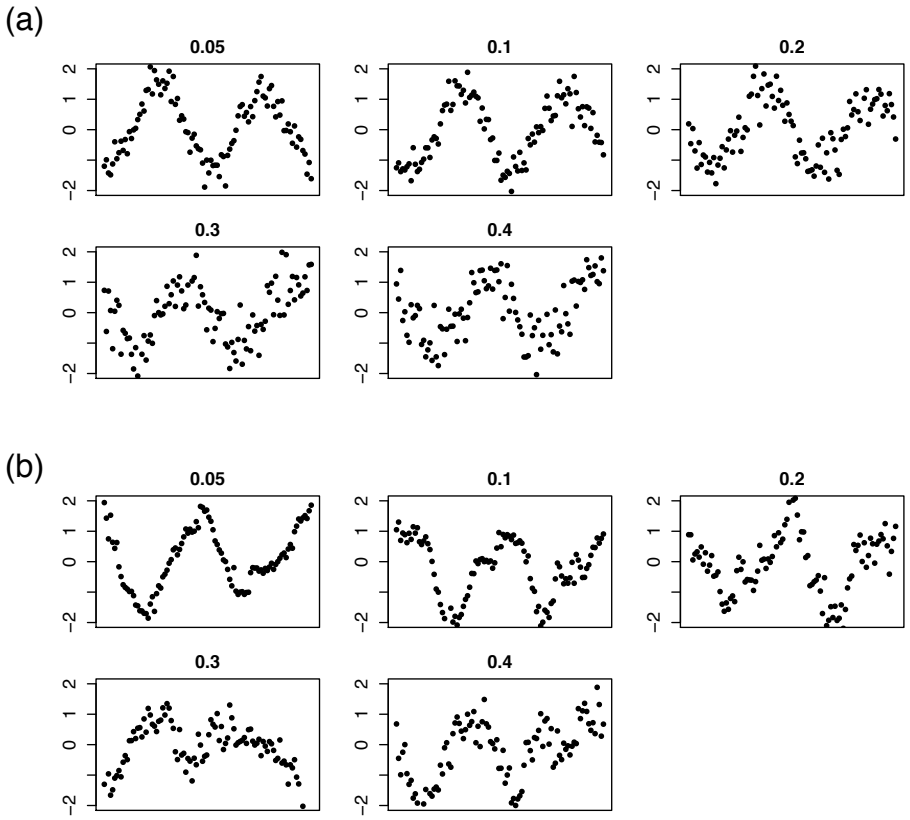


Supplementary Figure 3: Panel (a) shows scaled expression for the 151 genes identified by Oscope. Rows indicate genes and columns indicate samples. The samples are shown following the order recovered by ENI. Yellow and blue represent high and low values. Panel (b) is similar to (a), but the samples are shown following the original order over time. The genes are shown in the same order in panels (a) and (b). The x-axis of Panel (c) shows the top number of genes identified by Whitfield *et al.*, 2002 using known time course order. The y-axis shows the number of genes that were also in the 151 genes identified by Oscope. The $x = y$ diagonal line is shown in blue.

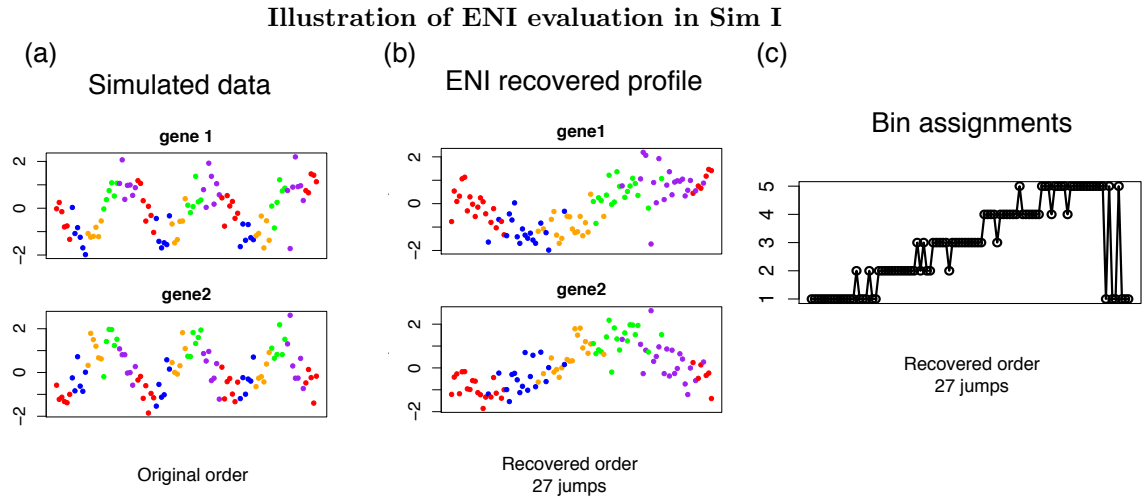
Example oscillatory genes with varying speeds in simulation studies



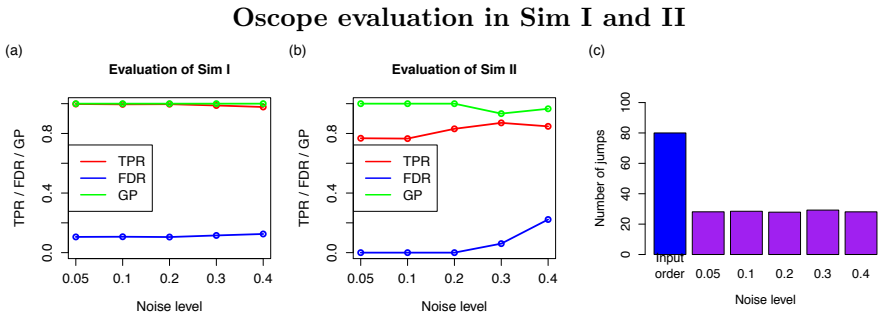
Supplementary Figure 4: Shown are example oscillatory genes in 3 frequency groups in **Sim I** (upper panels) and **Sim II** (lower panels). The x-axis shows original order along the simulated time series (note group 1 and 3 share the same order, while group 2 follows another order). The y-axis shows expression. σ_{str} is defined as 0.2 for these genes.



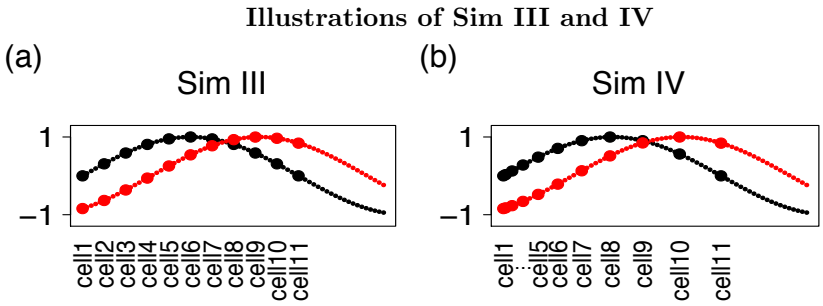
Supplementary Figure 5: (a) Shown are example oscillatory genes with varying noise levels under scenario **Sim I**. The x-axis shows the original orders and the y-axis shows the expression. σ_{str} varies from 0.05 to 0.4. (b) Shown are example oscillatory genes with varying noise levels under scenario **Sim II**.



Supplementary Figure 6: (a) Shown are 2 genes from one data set simulated following **Sim I** with $\sigma_{str} = 0.2$. Samples from different bins are shown in different colors. (b) Shown are the ENI recovered base cycle profiles of 2 genes in (a). Samples are colored by their bin assignment in their reference bin specification. (c) Shown are bin assignments of samples based on the ENI recovered order.

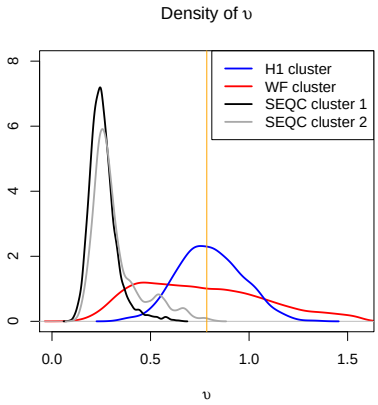


Supplementary Figure 7: (a) Shown are evaluation results from 5 simulation studies with varying noise levels under **Sim I**. The y-axis shows Oscope's TPR, FDR and GP averaged over 10 simulations within each group. The x-axis shows σ_{str} defined in each study. (b) Shown are evaluation results from 5 simulation studies with varying noise levels under **Sim II**. (c) Shown are ENI evaluation results from 5 simulation studies with varying noise levels under **Sim I**. We ran 10 repeated simulations for each study. The y-axis shows the average number of jumps based on Oscope's recovered order, across gene clusters and 10 repeated simulations. The x-axis shows σ_{str} defined in each study. The first bar shows the median number of jumps across all simulated data sets where median is taken over the random permuted input orders of 50 simulations.

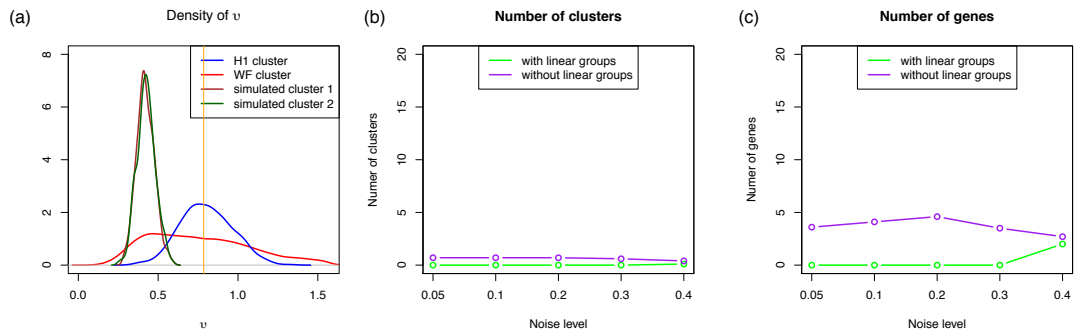


Supplementary Figure 8: Shown are illustrations of (a) **Sim III** and (b) **Sim IV**. In each of the simulations, we simulated two genes and 11 cells. Expression of the two genes are shown in black and red, respectively. The x-axis shows the oscillation time. The y-axis shows expression. Here φ_{g1} and ψ were set to 0 and 1, respectively. In **Sim III**, cells were evenly sampled. **Sim IV** is similar to **Sim III**, but the cells were not evenly sampled.

Density plots of phase-shift residuals on SEQC data

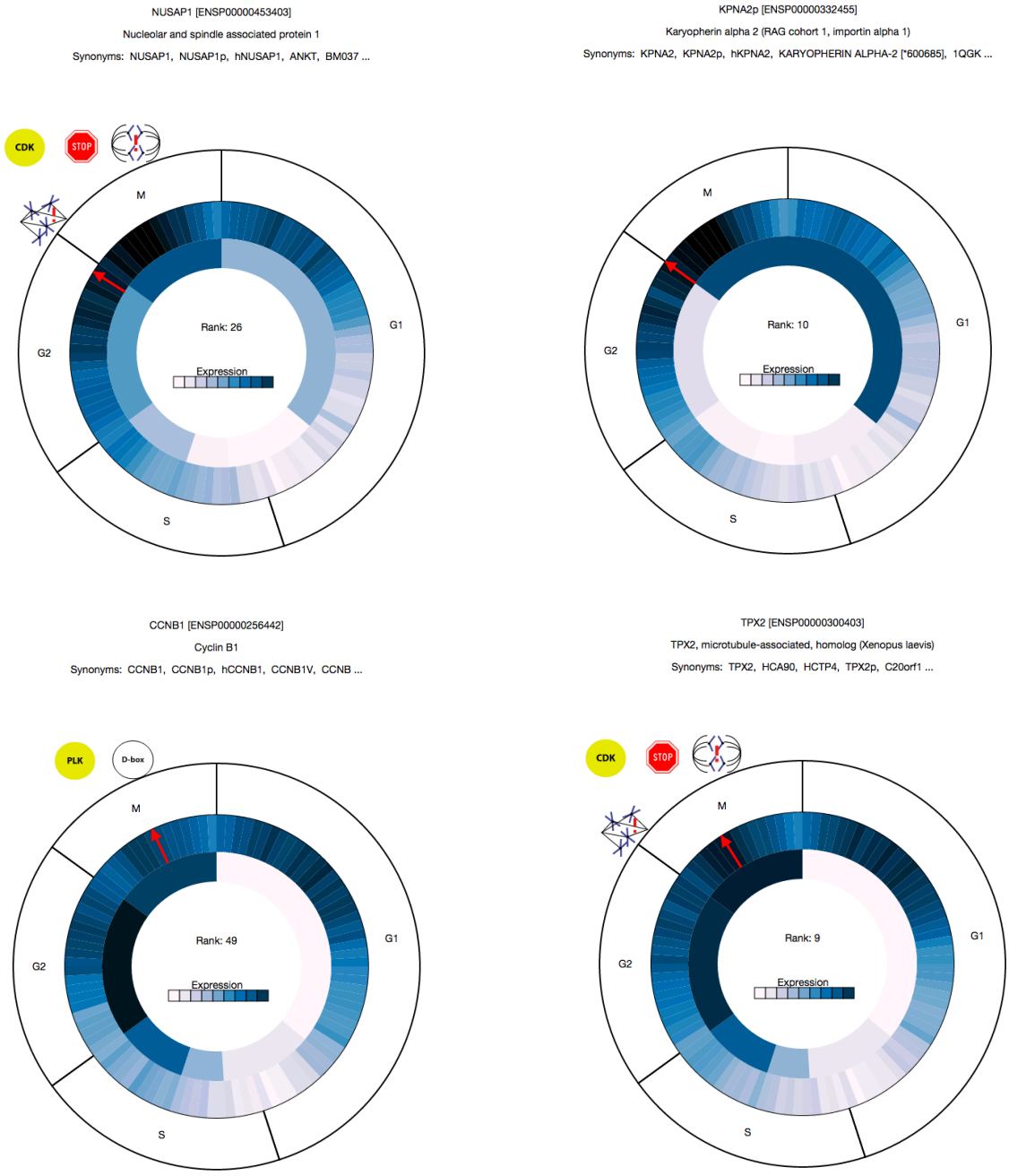


Supplementary Figure 9: Density plots of phase-shift residuals for clusters of genes identified by applying Oscope to three datasets. The phase-shift residuals for the cluster of 29 genes from H1 and the cluster of 151 genes from the Whitfield data are shown in blue and red, respectively. The phase-shift residuals from two clusters identified by the K-medoids algorithm (but filtered out in a later step of Oscope) from the SEQC data are shown in black and grey, respectively. For a cluster to be passed onto the ENI step in Oscope, the 90th quantile of the phase-shift residuals must exceed the $\pi/4$ threshold marked by an orange line.

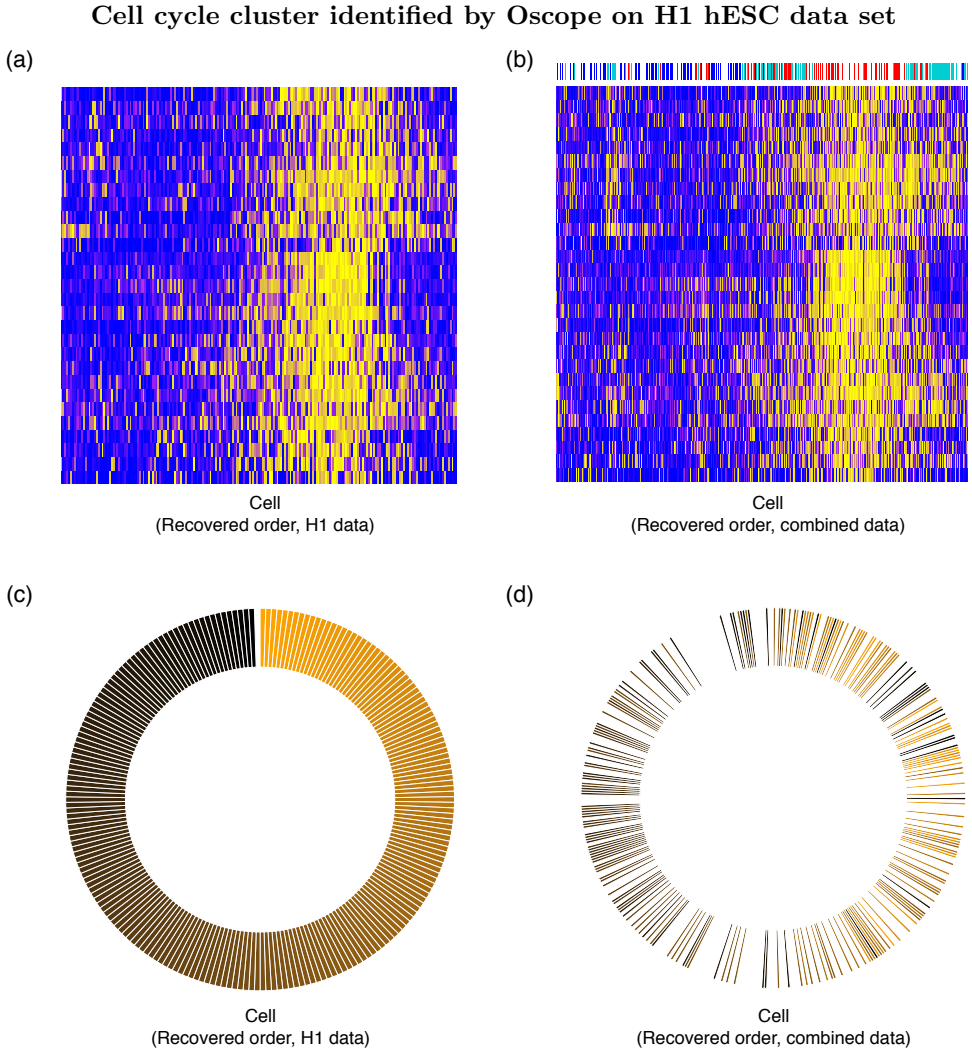


Supplementary Figure 10: (a) Density plots of phase-shift residuals for clusters of genes identified by applying Oscope to H1 data, Whitfield data, and a simulated data set with linear co-regulated gene groups. The phase-shift residuals for the cluster of 29 genes from H1 and the cluster of 151 genes from the Whitfield data are shown in blue and red, respectively. The phase-shift residuals from two clusters identified by the K-medoids algorithm (but filtered out later) from the simulated data are shown in brown and dark green, respectively. Panels (b) and (c) show average number of clusters and genes identified by Oscope in simulation studies described in Supplementary section Section 5.2. Simulations with and without linear co-regulated gene groups are shown in green and purple, respectively.

Cyclebase profiles of the 4 genes shown in Figure 2c-d

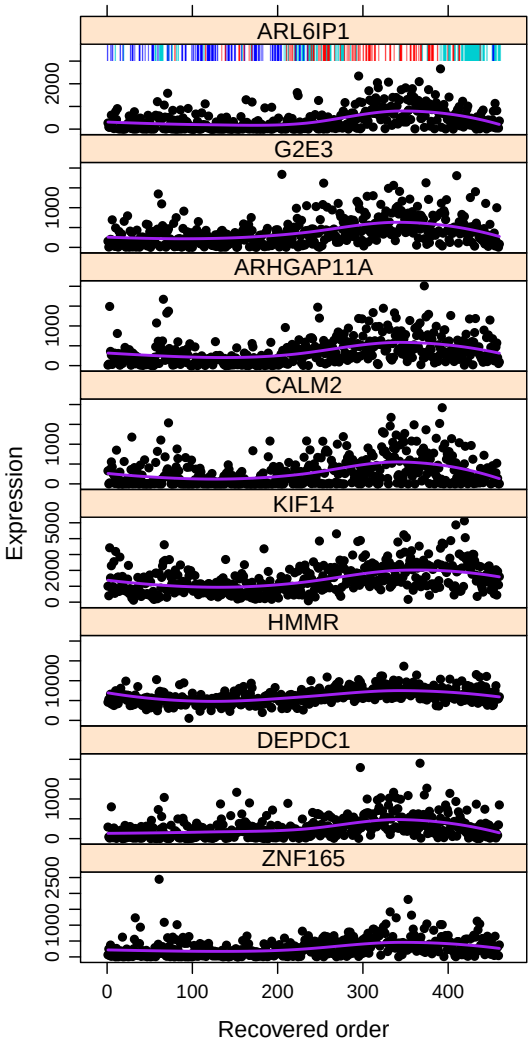


Supplementary Figure 11: Cyclebase profiles of the 4 genes (NUSAP1, KPNA2, CCNB1, and TPX2) shown in **Figure 2c-d**. The red arrow indicates Cyclebase estimated peak of expression in human. The Cyclebase estimates were generated based on previously published transcriptomics and/or proteomics time series data sets.

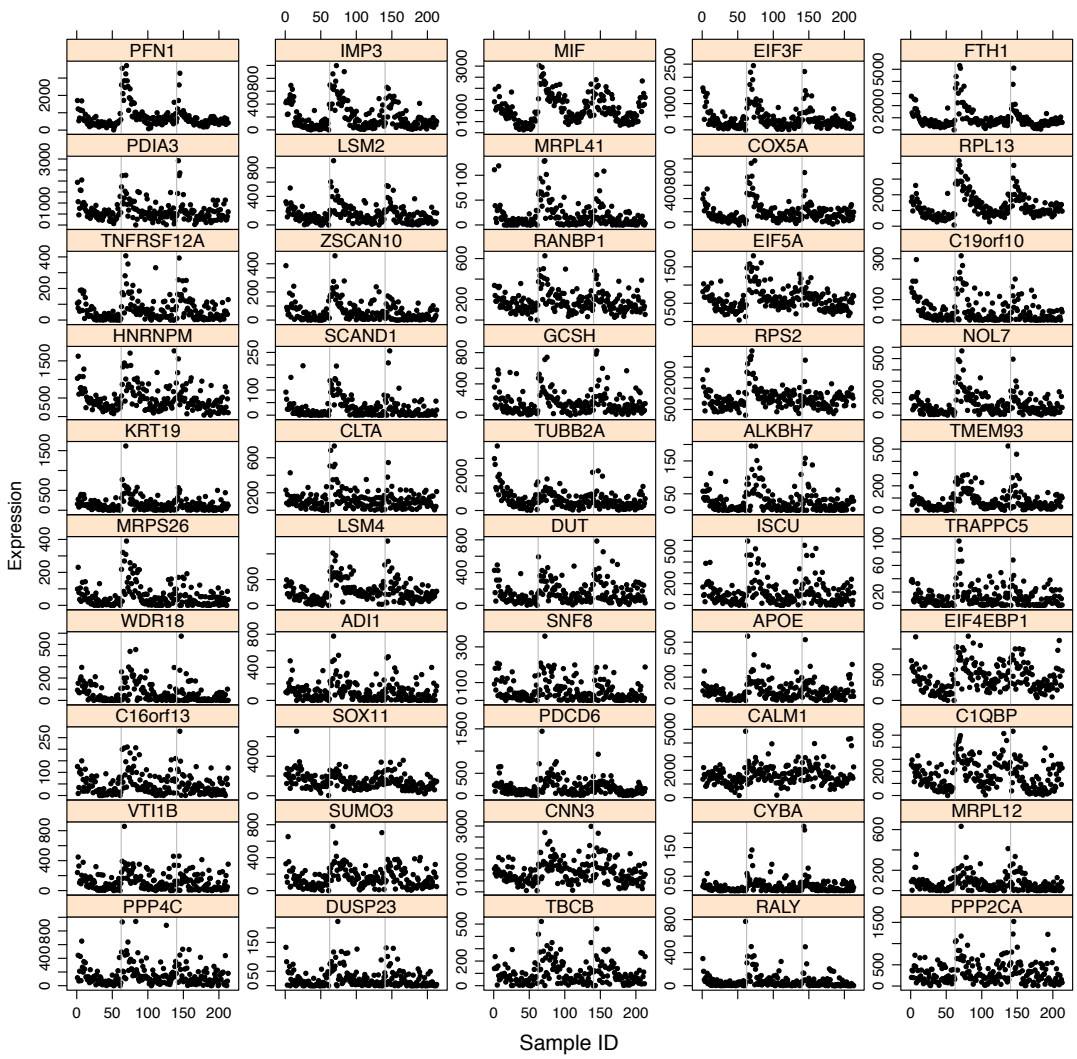


Supplementary Figure 12: Panel (a) shows scaled expression for the 29 genes identified by Oscope. Rows indicate genes and columns indicate samples. The samples are shown following the order recovered by ENI on the data set of 213 H1 hESCs. Yellow and blue represent high and low values. Panel (b) is similar (a), but showing results using combined data of 460 cells (similar to **Fig. 2d**). The genes are shown in the same order in panels (a) and (b). In panel (b), cells from S, G2/M and G1 are marked with blue, red or turquoise marks above the heatmap. Panel (c) shows the H1 hESCs following the recovered order on H1 data (the same order as in panel (a) and **Fig. 2c**. The cells are colored by their relative position on the base cycle. Panel (d) shows the H1 hESCs following the recovered order on the combined data (the same order as in panel (b) and **Fig. 2d**. Each H1 hESC is shown in the same color as in (c), and the H1-Fucci cells are shown in white.

Genes identified by Oscope that are not in the cell cycle GO category

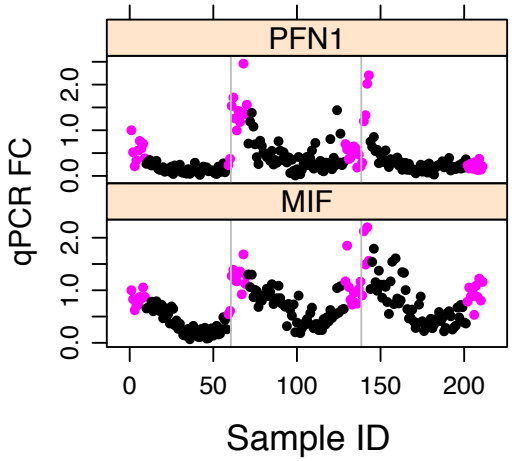


Supplementary Figure 13: Shown are the eight genes identified by Oscope that are not in the cell cycle GO category. Similar to **Figure 2d**, the x-axis shows the 460 cells following the recovered order on the combined data. Cells from S, G2/M and G1 are marked with blue, red or turquoise marks.



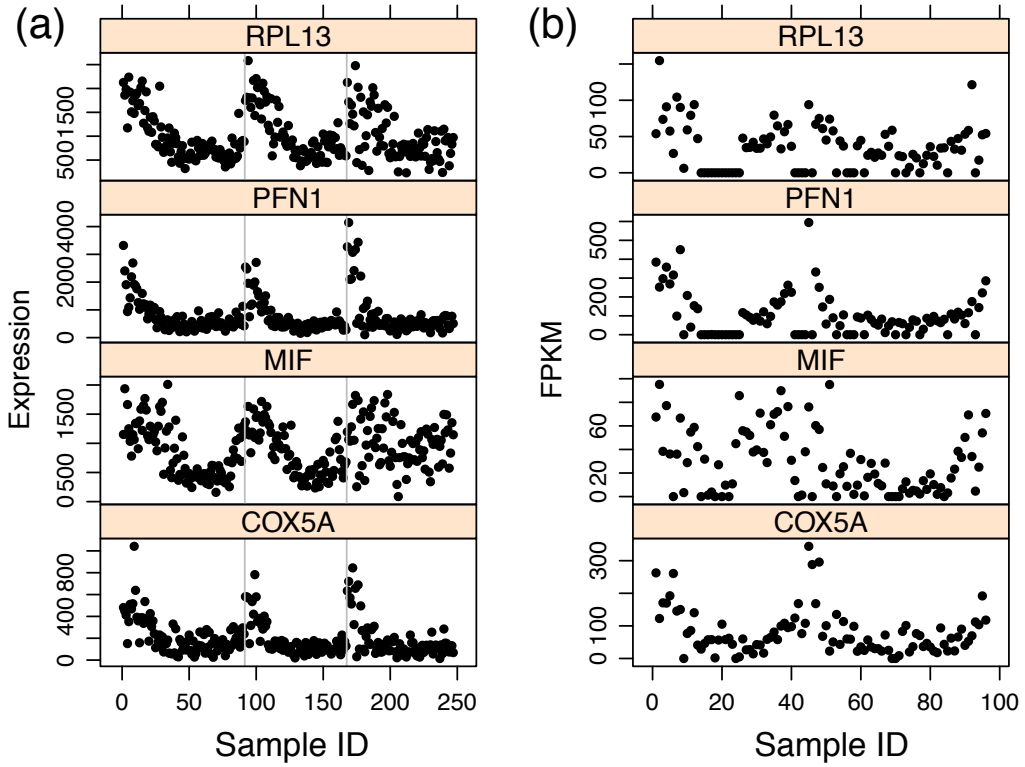
Supplementary Figure 14: Shown are the top 50 genes with ordering effects in the 213 H1 hESCs. The x-axis shows sample ID's and the y-axis shows normalized expression. Data from the three replicate H1 hESC experiments are separated by the gray lines.

qPCR validation of genes with ordering effects on H1 single cell cDNA



Supplementary Figure 15: Shown are qPCR results for PFN1 and MIF. The y-axis shows relative gene expression fold changes normalized to GAPDH in each single-cell cDNA. The relative expression value of single-cell sample #1 was set to 1 for comparison. The data from the three replicate H1 hESC experiments are separated by gray lines. Samples are shown following the C1 sample ID on the x-axis. Cells from the marked capture sites in **Figure 3a** are shown in magenta.

Example genes with ordering effects in additional public available scRNA-seq data sets



Supplementary Figure 16: Shown are 4 genes with ordering effects in 2 different data sets. The x-axis shows sample ID's and the y-axis shows normalized expression or FPKM. (a) H1-Fucci data, three experiments are separated by gray lines (each enriched for cells from G1, G2/M or S phase). (b) Data set from Wu *et al.*, 2014[1]. On the x-axis, cells are shown following the original sample order in the supplementary data set of Wu *et al.*. The y-axis shows the FPKM obtained from the supplementary data set.

Evaluation of Sim III and IV

	0	0.05	0.1	0.2	0.3	0.4
Sim III	0%	0%	0%	0.05%	0.58%	3.39%
Sim IV	0%	0%	0%	0.05%	0.70%	3.86%

Supplementary Table 1: Evaluation of **Sim III** and **IV**. The columns show results from 6 simulation studies with varying noise levels. For each simulation study, the average percentage is calculated across the 3969 simulations with varying φ_{g1} and ψ.

Section 1 Order recovery of multiple independent oscillatory groups

Supplementary Figure 1 shows two independent oscillatory gene groups. For simplicity, only one gene is shown for each group. This is different than **Figure 1**, where two genes from the same group are shown. We assume cells are from an unsynchronized population, and consequently the oscillation start time varies for different cells. For genes from two independent groups, the oscillation start time of these two genes also varies. For example, gene 1 might be a circadian gene while gene 2 is a somitogenesis gene. Cell 1 might enter the somitogenesis clock at the late state of circadian rhythm, while cell 2 could enter at an early state. Recall from **Figure 1** that the base cycle(s) can be recovered if we are able to sort cells by their oscillation time. As shown in **Supplementary Figure 1b**, sorting cells based on the oscillation time of gene 1 gives a different order than sorting by the oscillation time of gene 2. **Supplementary Figure 1c** shows that if two genes are from independent oscillatory groups, their expression scatter plot will no longer be ellipse shaped.

Section 2 Oscillation with varying speed or partial synchronization

Supplementary Figure 2 illustrates that the paired-sine model in Oscope is robust to cases where oscillations have varying speed or when partial synchronization (uneven sampling) occurs. The first two panels of **Supplementary Figure 2a** show expression of two co-regulated oscillatory genes over oscillation time. Both genes follow a sinusoid signal (with some phase shift) for part of their oscillation but have a slower progression in the interval between 20min and 120min. Thirty cells were collected simultaneously and the cell population is unsynchronized. The 30 cells are shown as dots in the first two panels. The third panel shows the co-regulation profile between these two genes. The elliptical shape is preserved even when the relationship between time and phase is not linear. Since the paired sine-model is independent of time, it does not require a specification regarding the relationship between time and phase. The first two panels

of **Supplementary Figure 2b** show expression of another two co-regulated oscillatory genes. The oscillation speed is constant over time. Thirty cells were collected simultaneously and a sub-group of the cells are synchronized at the beginning of the oscillation. The third panel shows the co-regulation profile between these two genes, where the elliptical shape is preserved as well. With respect to assumptions in the ENI algorithm, to avoid making assumptions regarding time or sampling, we used polynomial fitting instead of fitting a sinusoid signal.

Section 3 Case study results on Whitfield data

Supplementary Figures 3a-b show 151 genes plotted following (a) recovered order or (b) known time-series order. Panel (a) indicates that Oscope recovered smooth base cycle profiles for these genes. Panel (b) indicates that genes identified by Oscope show cyclic patterns in original time series experiments. Comparing panel (a) with (b), the results indicate that the Oscope recovered order is able to reveal the phase shifts across different genes.

Supplementary Figure 3c shows how the 151 genes distribute among the gold-standard list of 874 genes identified in Whitfield *et al.*, 2002. Oscope identified 47 out of the top 50 genes in the Whitfield list. Most of the genes that were not in the Oscope cluster have lower ranks in the Whitfiled list.

Section 4 Simulation studies

Section 4.1 Simulation set-up

We conducted two simulation scenarios to evaluate the Oscope pipeline. In **Sim I**, for a given oscillatory gene g in a cell s with oscillation time t_s, we simulated expression $X_{g,ts}$ as a sinusoid signal following $Sin(\omega_g t_s + \varphi_g) + \epsilon_{g,ts}$, in which ω_g is the angular frequency and φ_g is the starting phase of gene g; $\epsilon_{g,ts}$ is Gaussian noise with mean 0 and standard deviation σ_g, where σ_g varies as described below.

In **Sim II**, instead of a sinusoid signal, we simulated oscillatory genes based on profiles of cyclic genes identified in Whitfield *et al.*, 2002[2]. Here $X_{g,ts}$ is simulated as $\mu_{g,ts} + \epsilon_{g,ts}$, where $\mu_{g,ts}$ is obtained from imputed Whitfield data and $\epsilon_{g,ts}$ is defined similarly as in **Sim I**. A group

of genes with linear trends were also simulated as detailed below.

In both simulation studies, we simulated 1,000 genes and 100 cells. 90 out of the 1,000 genes were simulated as oscillators. The 90 oscillators were simulated in 3 frequency groups, each group contains 30 genes. In all simulations, we simulated genes in group 1 and 3 following the same order, while genes in group 2 following another order. In **Sim I**, the relative speeds of the 3 groups are proportional to 2:3:6. Example genes can be found in the upper panels of **Supplementary Figure 4**.

In **Sim II**, we implemented an imputation algorithm to extend 48 samples in Whitfield *et al.*, 2002[2] to a time series data with 100 cells. We took the expression profiles of the first 100 cyclic genes defined in the original paper as a kernel signal in our imputation. Recall in the Whitfield data, 48 samples were measured and oscillators have roughly 3 cycles over the time series. To simulate a gene in the fastest group, we randomly selected one of the top 100 genes, repeated the 48 measures twice to obtain $\mu_{g,1}, ..., \mu_{g,48}$ and $\mu_{g,51}, ..., \mu_{g,98}$. We then imputed $\mu_{g,49}, \mu_{g,50}, \mu_{g,99}$ and $\mu_{g,100}$ by taking average expression of $\mu_{g,48}$ and $\mu_{g,1}$ plus random noise. To simulate a gene in the group with median speed, we also randomly selected one gene from the top 100, then extended the 48 measures to 100 cells. We first obtained $\mu_{g,1}, \mu_{g,3}, ..., \mu_{g,99}$ by taking the 48 measures from Whitfield data (the first 2 measures were used twice to get $\mu_{g,97}$ and $\mu_{g,99}$). Then we generated imputed measures $\mu_{g,2}, \mu_{g,4}, ...$ The imputed measure was generated as averaged expression of its two adjacent cells plus random noise. For the slowest group, the first 33 samples from the original data of a selected gene were used and the imputation was conducted in a similar way. As a result, 3 groups of genes have approximately 2, 3 and 6 cycles, respectively. Example genes can be found in the lower panels of **Supplementary Figure 4**.

Within each frequency group, genes were further simulated with strong and weak signals. Half of the oscillatory genes were simulated as strong oscillators with $\sigma_g = \sigma_{\text{str}}$. The other half were simulated as weak oscillators with $\sigma_g = \sigma_{\text{wk}} = 2 * \sigma_{\text{str}}$. Starting phase φ_g varies in different genes within a frequency group. The remaining genes except the oscillators are called noise genes. Noise genes were simulated as random Gaussian noise. The noise level was adjusted to be comparable to the average noise signal among all oscillators.

In **Sim II**, we also simulated a group of 30 genes having linear relationship but are not oscillating. The kernel signal is evenly sampled from [-0.5, 0.5]. Half of the 30 genes were

simulated as kernel signal + Normal$(0, \sigma_{\text{str}})$ and the other half were simulated as kernel signal + Normal$(0, \sigma_{\text{wk}})$.

Section 4.2 Simulation evaluation

To evaluate Oscope, we conducted 5 simulation studies with varying noise levels for each simulation scenario. Specifically, the σ_{str} varies from 0.05 to 0.4 in 5 steps. Note the noise level in the weak oscillatory group is always defined as $\sigma_{\text{wk}} = 2 * \sigma_{\text{str}}$. For each simulation study, 10 simulations were generated. **Supplementary Figure 5a** shows example oscillators with varying noise levels under scenario **Sim I**. Similar trace plots for **Sim II** are shown in **Supplementary Figure 5b**.

We applied Oscope on each simulated data set. The top 10% of genes were selected from the paired-sine model and further clustered into oscillatory groups by the K-medoids algorithm. To evaluate the paired-sine model, we consider two summary statistics:

TPR: $\frac{\text{Num True Positive genes detected by Oscope}}{\text{Num genes simulated as oscillating}}$

FDR: $\frac{\text{Num False Positive genes detected by Oscope}}{\text{Num genes detected by Oscope}}$

To evaluate the K-medoids algorithm, we call a group a True Positive if more than half of the genes in the group were simulated as oscillating. We also consider:

GP (Group-wise precision): $\frac{\text{Num True Positive groups detected by Oscope}}{\text{Num groups detected by Oscope}}$

The ENI algorithm was further applied to gene groups defined in the K-medoids clustering step. To evaluate the ENI recovered order, we consider statistics defined as follows. Recall in **Sim I**, gene expression $X_{g,ts}$ is simulated following $Sin(\omega_g t_s + \varphi_g) + \epsilon_{g,ts}$, and consequently the samples were simulated with an angle between $[0, 2\pi]$. For a gene with angular frequency ω_g, the angle of a cell s at time t_s can be calculated from $(\omega_g t_s \bmod 2\pi)$, which is the same as the relative base cycle position. We evenly divide $0 - 2\pi$ into 5 intervals. For each frequency group, we split samples into 5 bins based on their angles. **Supplementary Figure 6a** shows bin specification of frequency group 1 on a simulated data set with $\sigma_{str} = 0.2$. Samples in 5 bins are shown in different colors. Note the bin specification is different across different speed groups. If the base cycle profile was reconstructed successfully based on ENI recovered order, the five bins should be separated clearly.

We then evaluate the recovered order by calculating the number of jumps. For a given order, a jump is called if the bin classification of a sample is different from the previous sample. Therefore, a perfectly recovered order would give very few jumps, whereas a poor recovery will give many jumps. For example, based on our input order from a random permutation, the median number of jumps across all simulated data sets is 80. **Supplementary Figure 6b** shows the 2 simulated genes from **Supplementary Figure 6a** following the recovered order. Each sample is marked with the color representing its bin assignment. The recovered order has 27 jumps. **Supplementary Figure 6c** shows step plots of the bin assignment of samples based on the recovered order. It indicates that most jumps happened at transition points.

Section 4.3 Simulation results

Supplementary Figures 7a-b show operating characteristics evaluating Oscope on simulation studies under scenarios **Sim I** and **Sim II**, respectively. Under scenario **Sim I** in which oscillators were simulated following a sinusoidal signal, Oscope has high GP/TPR and well controlled FDR in all cases. When oscillators do not follow a sinusoidal signal (**Sim II**), Oscope has decreased performance, as expected. However, operating characteristics still show good performance with GP/TPR above/around 0.8 and the FDR controlled below 0.3 even for high noise levels.

We also evaluated the ENI results on all **Sim I** data sets. **Supplementary Figure 7c** shows the number of jumps based on the recovered order. Based on the ENI recovered orders, the number of jumps are controlled below 30, whereas the median number of jumps across the random permuted input orders of 50 simulations is 80. In **Sim II**, expression is imputed from empirical data, and so ground-truth is not known. As a result, the true number of jumps cannot be evaluated.

Section 4.4 Evaluation of the polynomial regression model used in ENI

Recall that in the ENI module, Oscope tries to find the optimal order which minimizes the aggregated MSE of the sliding polynomial regression. To evaluate the performance of the polynomial regression on choosing the optimal order, we further conducted two simulations scenarios **Sim**

III and **Sim IV**. For a given cell s, expression of a gene pair is simulated as:

$$X_{g1,ts} = Sin(t_s + \varphi_{g1}) + \epsilon_{g1,ts}$$

$$X_{g2,ts} = Sin(t_s + \varphi_{g1} - \psi) + \epsilon_{g2,ts}$$

in which $s = 1, 2, ..., S$; ϵ's $\sim$ N$(0, \sigma)$, i.i.d; t_s indicates the oscillation time of cell s; ψ indicates the phase shift between the two genes; φ_{g1} indicates the starting phase of gene 1; and then the starting phase of gene 2 can be written as $\varphi_{g1} - \psi$. In **Sim III**, cells were evenly sampled regarding the oscillation time (see **Supplementary Fig. 8a** for an illustration): $t_s = (s-1) * \pi/10$. In **Sim IV**, cells were unevenly sampled (see **Supplementary Fig. 8b**): $t_s = (s-1)^2 * \pi/100$. To ensure the uniqueness of the optimal order, we simulated all cells from a single period. To scan all possible starting phases and phase shifts among genes, we first considered two genes. Also, to display design matrices, we first consider 11 cells.

In our case studies and previous simulation studies, we applied the ENI algorithm with sliding polynomial regression with 3 degrees of freedom. So we also consider the polynomial regression with 3 degrees of freedom here:

$$X_{g,s} \sim \beta_{g,0} + \beta_{g,1} s + \beta_{g,2} s^2 + \beta_{g,3} s^3$$

The best unbiased estimator of this model can be obtained from $(A^T A)^{-1} A^T X_g$. Therefore the minimal RSS based on the original order is:

$$R_g = ||X_g - A\hat{\beta}_g||^2 = ||X_g - A(A^T A)^{-1} A^T X_g||^2, \ g = 1, 2$$

in which $X_g = (X_{g,t1}, ..., X_{g,t11})^T$, $\beta_g = \beta_{g,0}, ..., \beta_{g,3}$; and A is the design matrix:

$$A = \begin{bmatrix} 1 & 1 & 1^2 & 1^3 \\ 1 & 2 & 2^2 & 2^3 \\ 1 & 3 & 3^2 & 3^3 \\ 1 & 4 & 4^2 & 4^3 \\ 1 & 5 & 5^2 & 5^3 \\ 1 & 6 & 6^2 & 6^3 \\ 1 & 7 & 7^2 & 7^3 \\ 1 & 8 & 8^2 & 8^3 \\ 1 & 9 & 9^2 & 9^3 \\ 1 & 10 & 10^2 & 10^3 \\ 1 & 11 & 11^2 & 11^3 \end{bmatrix}$$

Reordering of the cells can be viewed as reordering the rows of the design matrix. The reordering will induce an updated design matrix $\tilde{A} = MA$. For example, if we switch the second and the third cell, the matrix M can be written as:

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}_{11 \times 11}$$

The minimal RSS of the alternative order can be written as:

$$\tilde{R}_g = ||X_g - \tilde{A}\hat{\tilde{\beta}}_g||^2 = ||X_g - \tilde{A}(\tilde{A}^T\tilde{A})^{-1}\tilde{A}^T X_g||^2 = ||X_g - MA(A^TA)^{-1}A^TM^TX_g||^2,\ g = 1, 2$$

Define $F = \sum_g \tilde{R}_g - \sum_g R_g$. Since the ENI algorithm will select the order with smallest aggregated MSE, it will choose an incorrect order only if the alternative order gives a negative F.

Under scenarios **Sim III** and **Sim IV**, we conducted 6 simulation studies with varying σ's from [0, 0.05, 0.1, 0.2, 0.3, 0.4] as in **Sim I** and **Sim II**. For each study, we conducted 3,969 simulations with varying φ_{g1}'s and ψ's in 0 to 2π. In each simulation, 1,000 random permuted orders were considered.

For each of the simulation studies, we evaluate the percentage of negative F's (incorrect orders) among the 1,000 alternative orders. **Supplementary Table 1** shows the average percentage of negative F's in each simulation study. Results show that when the noise level is low ($\sigma < 0.2$), none of the alternative orders has negative F. The percentage increases when the noise level increases, but is controlled under 4% in all cases. These results indicate that the ENI algorithm will be able to select the correct order in most cases. The percentage of negative F's are similar in **Sim III** and **IV**, indicating that the ENI algorithm is robust to cases when the cells are not evenly sampled regarding the oscillation time.

We initially considered two genes so that all possible starting phases and phase shifts could be evaluated; and we initially considered 11 cells so that design matrices could be displayed. A similar study using **Sim III** and **Sim IV** with 20 genes and 50 cells gave similar results (all possible starting phases and phase shifts could not be evaluated).

Section 5 Evaluation of Oscope on data sets without oscillatory signals

To further evaluate the potential for false discoveries, we applied Oscope on one bulk RNA-seq data set which is not expected to contain any oscillatory signals. We also applied Oscope on simulated data sets containing no oscillators. In summary, Oscope did not identify any oscillatory

gene groups in the bulk RNA-seq data set, and detected very few genes in simulated data sets. Specifics follow.

Section 5.1 SEQC data set

To ensure a large enough sample size, we considered a bulk RNA-seq data set from the SEQC III project, which was generated in BGI. The sample was derived from Agilent's Universal Human Reference RNA (UHRR). The data set contains 80 technical replicates; therefore no oscillatory signal is expected. After applying Oscope on the SEQC data, no oscillatory genes were found. Specifically, using Oscope's default settings, the K-medoids algorithm gave 2 gene groups. However, neither group was passed to the ENI step because of the lack of within-cluster phase shift. Specifically, for a group of genes to be passed onto the ENI step, the 90th quantile of its $\upsilon_{gi,gj}$'s (phase-shift residuals) must exceed $\pi/4$. **Supplementary Figure 9** shows the phase-shift residuals for the SEQC clusters compared with the Whitfield and H1 clusters (where oscillators are detected). The phase-shift residuals in the 2 SEQC clusters are close to 0 compared with those in the other 2 clusters. As evident in the figure, the phase-shift residuals do not exceed the required $\pi/4$ threshold (orange line) and consequently no oscillatory genes are identified by Oscope.

Section 5.2 Simulated data sets without oscillatory signals

To further evaluate Oscope on simulated data, we conducted a simulation study similar to **Sim I** and **Sim II** in Section 6, with 1,000 genes and 100 cells, but with no oscillatory genes. To mimic the SEQC data, we simulated 2 groups of genes to be linearly related, each containing 50 genes. The other 900 genes were simulated as random noise. The simulations were conducted with varying noise levels, following the same settings as in **Sim II**. We repeated the simulation 10 times for each noise level. **Supplementary Figures 10b-c** show the average number of FD clusters and average number of FD genes identified by Oscope. Results indicate that Oscope identified very few genes in simulations with only linearly co-regulated gene groups.

Similar to **Supplementary Figure 9**, **Supplementary Figure 10a** shows the phase-shift residual densities. Instead of showing the SEQC data, shown are two clusters that were called by the K-medoids algorithm but were filtered out by the phase-shift residual criteria in one

simulation.

We also conducted a simulation without linear co-regulated gene groups. As above, we simulated 1,000 genes and 100 cells for each simulation. All 1,000 genes were simulated as noise. We repeated the simulation 10 times for each noise level. **Supplementary Figures 10b-c** also show the average number of FD clusters and average number of FD genes identified by Oscope in this simulation study. Results show that Oscope identified a very small number of genes in simulations of pure noise.

Section 6 Case study results on H1 hESC data

Section 6.1 The cell cycle gene set identified by Oscope

Supplementary Figure 11 shows Cyclebase[3] profiles of the 4 genes shown in **Figure 2c-d**. The red arrows indicate each gene's Cyclebase estimated time of expression peak in human. The Cyclebase estimates suggest that the peaks of NUSAP1 and KPNA2 are in late G2 phase and the peaks of CCNB1 and TPX2 are in the middle of M phase. These estimates agree with the phase shifts suggested by Oscope recovered profiles in **Figure 2c-d**.

Supplementary Figures 12a-b show 29 cell cycle related genes identified by Oscope on the H1 hESC data. Similar to **Figure 2c-d**, genes are plotted using recovered order on (a) data generated from H1 hESCs or (b) combined data set of H1 and H1-Fucci hESCs. Panel (b) shows that the order recovered by Oscope successfully separated the three cell cycle phases of the H1-Fucci cells. In addition, Oscope recovered smooth base cycle profiles for these genes on both H1 data and the combined data.

To further confirm that the recovered base cycle profiles on H1 data are consistent with the recovered profiles on the combined data, we compare the cell order of the 213 H1 hESCs in these two recoveries. **Supplementary Figure 12c** shows the H1 hESCs following the recovered order on H1 data (the same order as in **Fig. 2c** and **Supplementary Fig. 12a**). The cells are colored by their relative position on the base cycle. **Supplementary Figure 12d** shows the H1 hESCs following the recovered order on combined data (the same order as in **Fig. 2d** and **Supplementary Fig. 12b**). Each H1 hESC is shown in the same color as in **Supplementary**

Figure 12c, and the H1-Fucci cells are shown in white. The spearman correlation between these two orders is 0.52. To test the significance of the association between these two orders, we conducted an association test with 10^5 permutations. In each permutation, we calculated the correlation between the order shown in **Supplementary Figure 12d** and a randomly permuted order. None of the permuted orders gave an absolute correlation greater than 0.52 so the p-value of the association test is less than 10^{-5}. The results show that the recovered orders of H1 hESCs are consistent in two data sets and that the recovered base cycle profiles using the H1 data set are also associated with cell cycle.

Among the 29 genes, 8 are not included in the cell cycle GO category. **Supplementary Figure 13** shows the recovered base cycle profiles of these 8 genes on the combined data set. All 8 genes display oscillatory profiles along the cell cycle phases.

Section 7 Examine ordering effects in multiple data sets

Section 7.1 Top 50 genes with ordering effects in H1 hESCs

We used an ANOVA model to identify genes with potential ordering effects; 403 genes were identified with ANOVA p-value less than 0.005. **Supplementary Figure 14** shows the expression of the top 50 genes.

Section 7.2 qPCR validation of genes with ordering effects on H1 single cell cDNA

To further investigate at which stage the ordering effect was introduced, we performed qPCR analyses on the full-length single-cell cDNA libraries used in the three replicate H1 hESC experiments. These single-cell cDNA were harvested directly from the C1 chips with dilutions according to manufacture's protocols. Therefore, gene expression at this step is upstream of all the sequencing library preparations (for examples, the NexteraXT sample preparations), but downstream of the SMARTer chemistry (Clontech) performed inside the C1 IFC chips. **Supplementary Figure 15** shows qPCR results of relative PFN1 and MIF expression across single cells in the three replicate experiments. The ordering effect detected in the scRNA-seq data is

also present in this qPCR data.

Section 7.3 Example genes with ordering effects in additional public available scRNA-seq data sets

In addition to the H1 data set and the data set from Trapnell *et al.*, 2014[4], we further examined the ordering effects in the H1-Fucci data set and another publicly available data set generated by Fluidigm C1. The publicly available data set was obtained from the supplement of Wu *et al.*, 2014[1] where FPKM's of genes were provided. **Supplementary Figure 16** shows the same four genes as in **Figure 3b-c**. These four genes have ordering effects consistent across all data sets.

References

[1] Wu, A. R. *et al. Nature methods* **11**, 41–46 (2014).

[2] Whitfield, M. L. *et al. Molecular biology of the cell* **13**, 1977–2000 (2002).

[3] Santos, A., Wernersson, R. & Jensen, L. J. *Nucleic acids research* **43**, D1140–D1144 (2014).

[4] Trapnell, C. *et al. Nature biotechnology* (2014).