Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments

Ning Leng[1,2,6], Li-Fang Chu[2,6], Chris Barry[2], Yuan Li[1], Jeea Choi[1], Xiaomao Li[1], Peng Jiang[2], Ron M Stewart[2], James A Thomson[2–4] & Christina Kendziorski[5]

Oscillatory gene expression is fundamental to development, but technologies for monitoring expression oscillations are limited. We have developed a statistical approach called Oscope to identify and characterize the transcriptional dynamics of oscillating genes in single-cell RNA-seq data from an unsynchronized cell population. Applying Oscope to a number of data sets, we demonstrated its utility and also identified a potential artifact in the Fluidigm C1 platform.

Oscillations in gene expression play a critical role in many biological processes including somitogenesis, limb development and progenitor cell maintenance[1]. However, even for well-known oscillatory systems such as the cell cycle, expression characteristics such as the peak phase of genes have not been thoroughly studied in all cell types owing to technological limitations. Recent advances in live-cell imaging have improved the sensitivity and specificity with which continuous measurements can be made within a single cell[2], but constraints associated with reporters and detection channels limit monitoring to relatively few genes in any given experiment. RNA microarray or RNA-seq time-series experiments are often conducted in order to study transcriptional oscillations on a genome-wide scale[3], but heterogeneity in gene-specific frequency and phase make it difficult to identify an optimal sampling rate. These methods also require large quantities of synchronized starting material and average expression over thousands of cells, which may miss or even misrepresent[4] oscillations. Cell synchronization, when possible, attenuates a number of these problems for known oscillatory systems (typically the cell cycle) but can dramatically alter the transcriptional dynamics of others, and it does not facilitate *de novo* discovery.

Single-cell RNA-seq (scRNA-seq) has the potential to capture a more precise (not averaged) representation of oscillation dynamics genome wide in populations of single cells as well as to unmask oscillations that are missed in bulk expression experiments. However, continuous monitoring within a cell is not possible, and high-resolution scRNA-seq time-series experiments in distinct cells are prohibitive given the time required for sample preparation and sequencing. Even when scRNA-seq time-series experiments become feasible, challenges associated with rate heterogeneity, sampling and synchronization will remain. In addition, when tissue samples are studied, synchronization is not possible for most oscillatory systems such as cell cycle.

Computational algorithms have been developed to address some of these challenges in both microarray[5–7] and scRNA-seq studies[4], but none focuses on identifying oscillating genes. Most are based on the recognition that different samples represent distinct states in a system, such as time points along a continuum or progression toward an end point. By obtaining multiple samples at a single[5–7] or a few[4] time points and computationally reconstructing an appropriate order, temporal or other meaningful dynamics can be resolved. A key assumption that enables ordering is that genes do not change direction very often and thus samples with similar transcriptional profiles should be close in order. Oscillating genes pose challenges for these approaches because genes following the same oscillatory process need not have similar transcriptional profiles. Two genes with an identical frequency that are phase shifted, for example, will have little similarity (**Fig. 1a**).

We have developed a statistical approach called Oscope, which is freely available, to identify oscillating genes in static, unsynchronized scRNA-seq experiments (**Fig. 1** and **Supplementary Software**). Like previous algorithms, Oscope capitalizes on the fact that cells from an unsynchronized population represent distinct states in a system. However, unlike previous approaches, ours does not attempt to construct a linear order on the basis of minimizing change among adjacent samples. Rather, Oscope utilizes co-regulation information among oscillators to identify groups of putative oscillating genes and then reconstructs the cyclic order of samples for each group, defined as the order that specifies each sample's position within one cycle of the oscillation (referred to as a base cycle). The reconstructed order aims to recover gene-specific cyclic profiles defined by the group's base cycle, allowing for phase shifts between different genes (Online Methods). Notably, for different groups of genes following independent oscillatory processes and/or having distinct frequencies, the cyclic orders of cells need not be the same (see **Supplementary Fig. 1**).

A single cell can be thought to oscillate through cell states defined, for simplicity, by oscillations in just two genes (**Fig. 1a**). In a typical scRNA-seq experiment, unsynchronized cells in a variety of different states are collected at the same calendar time T (**Fig. 1b**). If it were possible to sort cells by the oscillation times of

[1]Department of Statistics, University of Wisconsin–Madison, Madison, Wisconsin, USA. [2]Morgridge Institute for Research, Madison, Wisconsin, USA. [3]Department of Cell and Regenerative Biology, University of Wisconsin–Madison, Madison, Wisconsin, USA. [4]Department of Molecular, Cellular, and Developmental Biology, University of California, Santa Barbara, California, USA. [5]Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, Wisconsin, USA. [6]These authors contributed equally to this work. Correspondence should be addressed to C.K. (kendzior@biostat.wisc.edu).

RECEIVED 18 DECEMBER 2014; ACCEPTED 5 JUNE 2015; PUBLISHED ONLINE 24 AUGUST 2015; DOI:10.1038/NMETH.3549

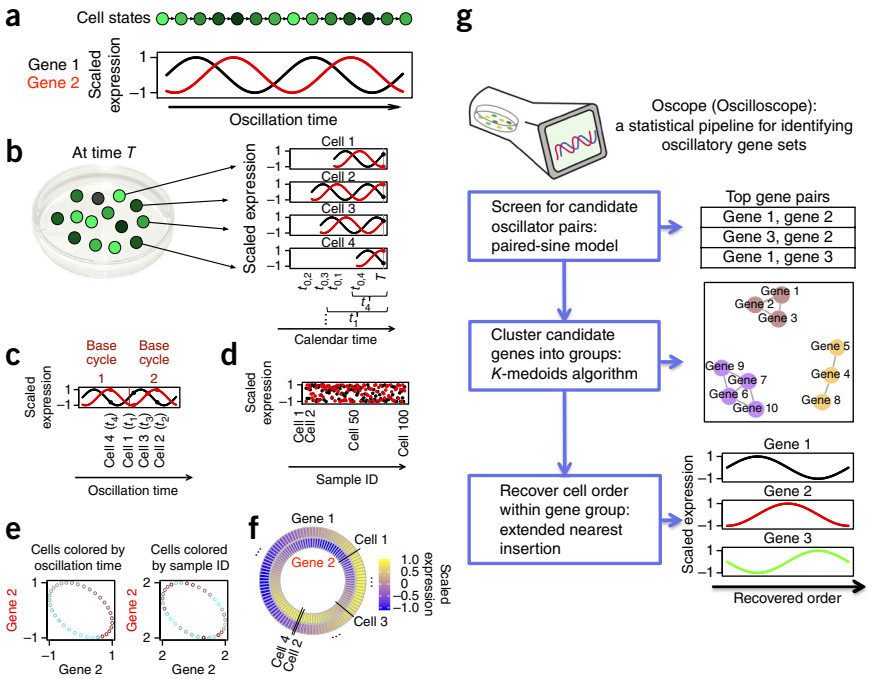

Figure 1 | Overview of Oscope. (**a**) An oscillating gene group with two genes and corresponding cell state. (**b**) In an unsynchronized scRNA-seq experiment, mRNA is collected at time T from cells in varying states. $t_{0,i}$ and t_i show cell i's oscillation start time and oscillation time, respectively. (**c**) The same genes and cells as in **b**, where cells are ordered by the genes' oscillation times. (**d**) Expression for 100 unsynchronized cells. (**e**) Scatter plots of gene 1 vs. gene 2 scaled expression, which are independent of order. Cells are colored from cyan to brown following the x axes of **c** and **d**, respectively. (**f**) Results of base-cycle reconstruction for the 100 cells shown in **d**. (**g**) Flow chart of the Oscope pipeline (see Online Methods).

genes, defined as the amount of calendar time the cell has been oscillating before collection time T, identifying oscillating genes and characterizing their dynamics would be straightforward (**Fig. 1c**). However, oscillation time is unobserved in an scRNA-seq experiment. With this type of snapshot data, the expression of oscillating genes is indistinguishable from random noise (**Fig. 1d**); therefore, existing methods[8,9] for identifying cyclic features do not apply.

Recognizing that a scatter plot of expression values for genes oscillating with similar frequency will form an ellipse independent of order (**Fig. 1e**), Oscope fits a two-dimensional sinusoidal function to all gene pairs and chooses those with the best fits. Note that the elliptical shape is preserved when the oscillation has varying speed or is partially synchronized between genes (see **Supplementary Fig. 2**). Once candidate genes are identified, the K-medoids algorithm is applied to cluster genes into groups with similar frequencies but possibly different phases. Then, for each group, Oscope recovers the cyclic order that places cells by their position within one cycle of the oscillatory process underlying the group. Given static data, it is not possible to reconstruct multiple cycles of an oscillatory process because the dynamics of late cycles are identical to those of earlier cycles, by definition. For example, the gene expression values in cells 2 and 4 in **Figure 1b** are identical even though cell 2 has passed through a full cycle but cell 4 has not. Here we define the base cycle as the minimal cycle that is repeated in an oscillatory process (an example is shown in **Fig. 1c**). Oscope uses an extended 'nearest-insertion' algorithm to order cells with respect to their position in a base cycle without specifying a direction of time (**Fig. 1f**).

The nearest-insertion algorithm[10] was developed to address the traveling salesman problem: given pairwise distances between cities, the algorithm provides a computationally efficient way to order travel to all cities so that overall distance is minimized. We extended the nearest-insertion algorithm

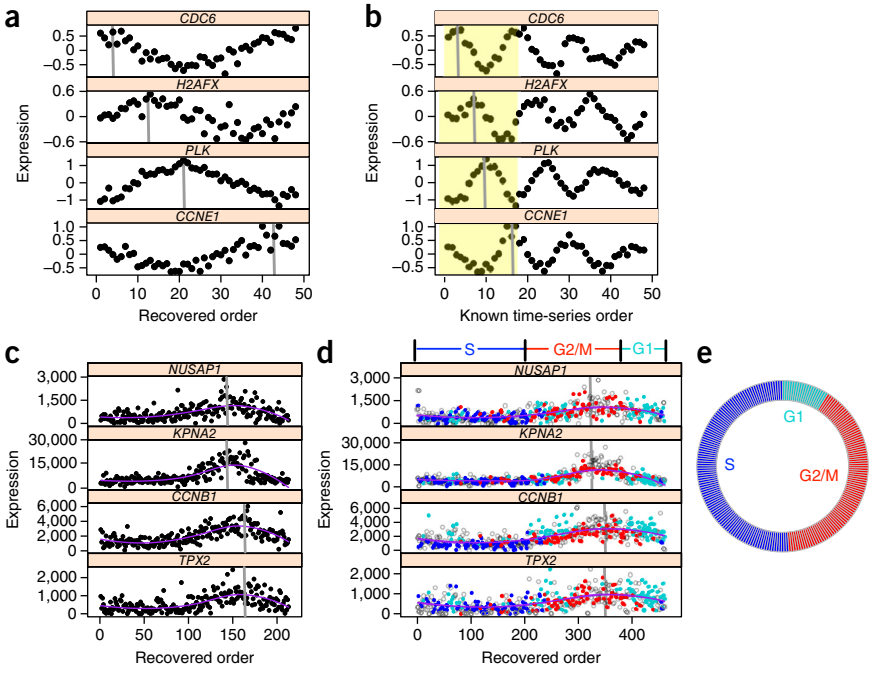

Figure 2 | Oscope uncovers oscillatory signals in transcriptional profiles. (**a**) Four genes in the time-series data from Whitfield *et al.*[8] with profiles ordered by Oscope; the peak of the base cycle is marked in gray. (**b**) The same four genes as in **a** following the known order over time with the peak of the first base cycle (shown in yellow) marked in gray. (**c**) Oscope-recovered profiles of four genes from a 29-gene group identified by Oscope using scRNA-seq data from 213 unlabeled hESCs. (**d**) The same four genes as in **c** ordered using 460 cells (213 unlabeled and 247 H1-FUCCI cells are shown as open and filled circles, respectively). FUCCI labels (ignored before applying Oscope) are shown in different colors for the 247 cells. Phase boundaries defined by the reconstructed order are shown above the plots. (**e**) The proportion of unlabeled cells that fall into each phase defined by the boundaries in **d**.

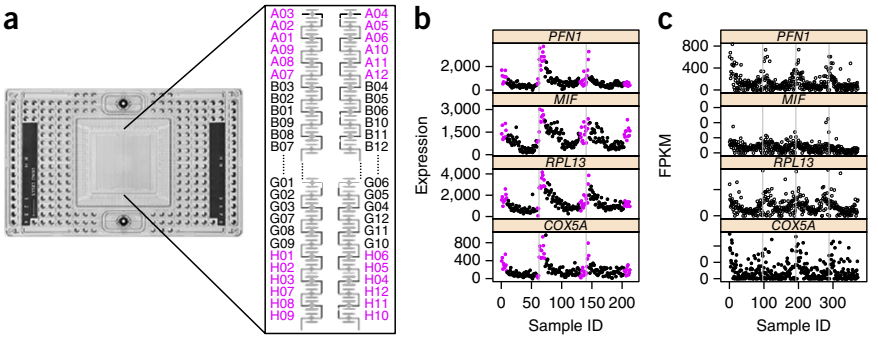

Figure 3 | Oscope uncovers dynamic signals of technical origin in scRNA-seq data sets. (**a**) Default plate output ID layouts of the capture sites on the C1 chip. The capture sites' corresponding plate output IDs are labeled following the recommendation by the manufacturer user guide. (**b**) Expression of four genes with potential ordering effects. Cells are ordered by the C1 plate output ID (A01–A12, B01–B12, …, H01–H12). Cells from the colored capture sites in **a** are also shown in magenta. Three replicate hESC experiments are separated by gray lines. (**c**) The same four genes for a different data set[4] (ordered following the cell order listed in supplementary data for ref. 4). The four experiments are separated by gray lines. The *y* axes are limited to the 98th quantile of gene-specific FPKMs (fragments per kilobase of exon per million mapped reads) for better visualization.

to order cells within an oscillatory gene group so that distance between each gene's expression and its gene-specific base-cycle profile is minimized on average over all genes in the group. Once the order for each group is recovered, subsequent algorithms developed for time-course analysis (for example, Fourier transformation, spline fitting, etc.) can be applied to estimate phase or further characterize oscillations, if desired.

To evaluate the ability of Oscope to identify oscillating groups of genes and reconstruct the cyclic order underlying their base cycles, we applied it to a bulk RNA-seq time-series study of oscillating genes[8] after permuting the sample order. The top group identified by Oscope had 151 genes, 116 of which were validated as oscillating in earlier work[8] (see **Fig. 2a,b** for examples). Oscope successfully recovered the base-cycle profile of each gene and correctly inferred phase shifts (**Supplementary Fig. 3** shows all 151 genes). The results of simulation studies and additional case studies provide further insights into the operating characteristics of the approach (**Supplementary Note**, **Supplementary Figs. 4–10** and **Supplementary Table 1**).

To further evaluate Oscope on scRNA-seq data, we analyzed profiles of single undifferentiated human embryonic stem cells (hESCs)[11]. We applied Oscope to three replicate scRNA-seq experiments on H1 hESCs (n = 213). One of the top groups identified by the K-medoids algorithm in Oscope contained 29 genes (**Supplementary Table 2**), 21 of which are annotated as belonging to the Gene Ontology "Cell Cycle" biological process (GO:0007049). The reconstructed base cycle is characterized by peaked expression of genes known to be involved in G2 phase progression (for example, *NUSAP1* and *KPNA2*) and M phase progression (for example, *CCNB1* and *TPX2*)[12] (**Fig. 2c** and **Supplementary Fig. 11**). To confirm whether the recovered profiles were associated with cell-cycle phasing, we performed additional scRNA-seq experiments (n = 247) on H1 hESCs harboring the fluorescent ubiquitination-based cell-cycle indicators[13] reporter (H1-FUCCI, see Online Methods) in which cells were identified as being in G1, S or G2/M phase. We combined the H1 and H1-FUCCI data sets and applied Oscope. The reconstructed order using the same 29 genes largely recapitulates the three phases of the cell cycle (**Fig. 2d** and **Supplementary Fig. 12**). The phase boundaries defined by the reconstructed order classified 72% of H1-FUCCI hESCs into the correct phase. Because the H1-FUCCI data set does not provide an unbiased estimate of the number of cells in each phase, we classified the unlabeled H1 hESCs by the phase boundaries and estimated the proportion in each phase. The proportion of unlabeled H1 cells in each phase is consistent with the notion of a shortened G1 phase in undifferentiated hESCs[14] (**Fig. 2e**). Out of the eight genes that were not annotated as belonging to the cell-cycle pathway, six of them have been shown to be associated with cell cycle in a previous publication[12]. All eight genes, including the two less well-characterized oscillatory genes *CALM2* and *ZNF165*, showed cell cycle–related base-cycle profiles (**Supplementary Fig. 13**).

A second group of top genes identified by Oscope showed an oscillatory pattern related to the capture site and output well positions on the Fluidigm C1 chip (**Supplementary Table 3**). In particular, these genes all had increased expression in cells captured in sites with small or large plate output IDs, across all three replicate hESC scRNA-seq experiments. The capture sites involving increased gene expression were physically located close to each other on the chip (**Fig. 3a**). To examine this potential artifact, we developed an analysis of variance (ANOVA)-based artificial trend detection algorithm (Online Methods) and applied the algorithm on the combined data from the three H1 experiments. We found that 403 genes showed strong artificial trends (**Supplementary Fig. 14** and **Supplementary Table 4**) consistently across experiments (**Fig. 3b**). To further investigate the artifact and to rule out biases that may be due to sequencing, we estimated expression via real-time quantitative PCR (qPCR) on select genes (**Supplementary Fig. 15**) and found the trend already present in the full-length single-cell cDNA libraries. We also saw this trend in publicly available data sets from other labs using various cell types (**Fig. 3c** and **Supplementary Fig. 16**).

The scRNA-seq technology offers an unprecedented ability to take snapshots of genome-wide transcription in single cells, but it is not amenable to longitudinal studies that monitor changes in individual cells *in situ*. Oscope allows investigators to identify and characterize oscillating gene groups and infer a gene's oscillation phase. Applications in a number of settings should improve our understanding of known oscillators as well as facilitate the discovery of new ones. Furthermore, adjusting for oscillators using the characterization provided by Oscope should increase the power to investigate other signals associated with differentiation and/or subpopulations[15].

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. NCBI Gene Expression Omnibus: GSE64016.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This work was supported by US National Institutes of Health grants GM102756, 4UH3TR000506 and 5U01HL099773, the Charlotte Geyer Foundation and the Morgridge Institute for Research. N.L. was supported by the Shapiro Fellowship. C.B. was supported by the Canadian Institutes of Health Research Banting Postdoctoral Fellowship. We thank M. Probasco and N. Propson for their assistance in sorting cells by FACS and J. Bolin, A. Elwell and B.K. Nguyen for the preparation and sequencing of the RNA-seq samples. We thank A. Gitter, K. Korthauer and R. Bacher for comments that helped improve the manuscript.

AUTHOR CONTRIBUTIONS

N.L., L.-F.C., R.M.S., J.A.T. and C.K. designed research, analyzed data and wrote the manuscript; C.B. generated the H1-FUCCI cell line; Y.L., J.C. and X.L. contributed to the simulation studies; and P.J. performed RNA-seq read mapping, quantification and quality control.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Aulehla, A. & Pourquié, O. *Curr. Opin. Cell Biol.* **20**, 632–637 (2008).
2. Shin, I. *et al. Nucleic Acids Res.* **42**, e90 (2014).
3. Bar-Joseph, Z., Gitter, A. & Simon, I. *Nat. Rev. Genet.* **13**, 552–564 (2012).
4. Trapnell, C. *et al. Nat. Biotechnol.* **32**, 381–386 (2014).
5. Qiu, P., Gentles, A.J. & Plevritis, S.K. *PLoS Comput. Biol.* **7**, e1001123 (2011).
6. Gupta, A. & Bar-Joseph, Z. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **5**, 172–182 (2008).
7. Magwene, P.M., Lizardi, P. & Kim, J. *Bioinformatics* **19**, 842–850 (2003).
8. Whitfield, M.L. *et al. Mol. Biol. Cell* **13**, 1977–2000 (2002).
9. de Lichtenberg, U. *et al. Bioinformatics* **21**, 1164–1171 (2005).
10. Rosenkrantz, D.J., Stearns, R.E. & Lewis, P.M. II. *SIAM J. Comput.* **6**, 563–581 (1977).
11. Thomson, J.A. *et al. Science* **282**, 1145–1147 (1998).
12. Santos, A., Wernersson, R. & Jensen, L.J. *Nucleic Acids Res.* **43**, D1140–D1144 (2015).
13. Sakaue-Sawano, A. *et al. Cell* **132**, 487–498 (2008).
14. Becker, K.A. *et al. J. Cell. Physiol.* **209**, 883–893 (2006).
15. Buettner, F. *et al. Nat. Biotechnol.* **33**, 155–160 (2015).

ONLINE METHODS

Oscope: paired-sine model. An oscillatory gene group is a group of genes having the same frequency with phase shifts that may vary among pairs but are preserved across cells. For example, if $\psi_{gi,gj,s}$ denotes the phase shift between genes *gi* and *gj* in cell *s*, then $\psi_{gi,gj,s}$ need not equal $\psi_{gi,gk,s}$, but $\psi_{gi,gj,1} = \psi_{gi,gj,2} = \psi_{gi,gj,S}$. Oscillation time is the difference between cell collection time *T* and the start of oscillation.

For a pair of genes *g*1 and *g*2, denote the matched gene expression (rescaled to [−1, 1]) in *S* cells as $(X_{g1,1}, X_{g2,1})$, $(X_{g1,2}, X_{g2,2})$, …, $(X_{g1,S}, X_{g2,S})$. If the two genes follow a sinusoidal process with a phase shift, then the following equations hold for each cell *s* in 1, 2, …, *S*: $X_{g1,s} = \sin(t_s + \varphi_{g1})$ and $X_{g2,s} = \sin(t_s + \varphi_{g1} + \psi_{g1,g2})$, where t_s indicates oscillation time of cell *s*, φ_{g1} indicates the starting phase of gene 1, and $\psi_{g1,g2}$ indicates the phase shift between the two genes where the subscript *s* is dropped because $\psi_{g1,g2}$ is assumed common to all cells.

By trigonometric identities

$$X_{g2,s} = \sin\left(t_s + \varphi_{g1}\right)\cos\left(\psi_{g1,g2}\right) + \cos\left(t_s + \varphi_{g1}\right)\sin\left(\psi_{g1,g2}\right)$$
$$= X_{g1,s}\cos\left(\psi_{g1,g2}\right) \pm \sqrt{1 - X_{g1,s}^2}\,\sin\left(\psi_{g1,g2}\right)$$

Given this, the following equation holds for any cell:

$$X_{g1,s}^2 + X_{g2,s}^2 - 2X_{g1,s}X_{g2,s}\cos\left(\psi_{g1,g2}\right) - \sin^2\left(\psi_{g1,g2}\right) = 0$$

and there exists an optimal $\psi_{g1,g2}$ for which the error term $\varepsilon_{g1,g2}^2$ is 0, where

$$\varepsilon_{g1,g2}^2 = \sum_s \left[X_{g1,s}^2 + X_{g2,s}^2 - 2X_{g1,s}X_{g2,s}\cos\left(\psi_{g1,g2}\right) - \sin^2\left(\psi_{g1,g2}\right) \right]^2$$

To search for gene pairs with associated dynamic changes, Oscope linearly rescales gene-specific gene expression measurements to range between −1 and 1, and estimates the optimal $\psi_{gj,gj}$ for all gene pairs (gene *i*, gene *j*) defined as that which minimizes $\varepsilon_{gi,gj}^2$. With this metric, gene pairs are rank ordered by $-\log_{10}(\varepsilon_{gi,gj}^2)$; and candidate oscillatory genes are those genes in the top gene pairs (Oscope's default is the top 5%; this threshold may be changed by users on the basis of the empirical distribution of the $\varepsilon_{gi,gj}^2$ values).

Oscope: *K*-medoids clustering. To cluster the candidate oscillatory genes detected from the paired-sine model into distinct groups, we use the *K*-medoids algorithm with $\varepsilon^2{}_{gi,gj}$ as the dissimilarity metric. With this metric, gene pairs with small $\varepsilon^2{}_{gi,gj}$ values are more likely to be clustered together. The optimal *K* is picked by maximizing the silhouette distance. Only groups having within-group phase differences are further considered in order recovery to avoid detecting gene groups with a purely linear relationship. Specifically, for any pair of genes *gi*, *gj* within a group, we define the phase-shift residual as $v_{gi,gj} = \min((\pi - \eta_{gi,gj}), \eta_{gi,gj})$, in which $\eta_{gi,gj} = (\psi_{gi,gj} \bmod \pi)$. Oscope's default takes groups whose 90th quantile of $v_{gi,gj}$ values is greater than $\pi/4$ for further order recovery.

Oscope: extended nearest insertion. We developed an extended nearest-insertion (ENI) algorithm to recover the cyclic order for each oscillatory group defined in the *K*-medoids clustering step. Cells are ordered cyclically according to their position within one cycle of the oscillation, referred to as a base cycle. The ENI starts with three randomly selected cells and forms a loop (undirected graph). A fourth cell is chosen at random and inserted into the three cell-cell gaps on the loop. This forms three candidate orders. We evaluate each order using the aggregated mean squared error (MSE) of a sliding polynomial regression (SPR). For a given order, SPR is fitted to the expression of each gene. To capture cyclic features of the data, SPR fits *m* polynomial regression models starting with *m* evenly distributed points on the loop. The largest MSE among the *m* models is defined as the MSE of the SPR for this gene. For each order, the aggregated MSE of an oscillatory gene group is defined as the summation of the MSEs among all genes. The optimal order of the first four cells is then selected as the one that minimizes the aggregated MSE. This process is repeated to insert the fifth cell and so on, until all cells are in the loop. A 2-opt algorithm is then applied to avoid finding local maxima.

Whitfield data and statistical analysis. Microarray gene expression data were downloaded from http://genome-www.stanford.edu/Human-CellCycle/HeLa/. In total, five experiments were available at this site from Whitfield *et al.*[8]; experiment 3 was used here as it has the largest sample size. For this experiment, double thymidine block was used to synchronize HeLa cells, and expression was profiled for 9,559 genes at 48 time points following synchronization. Gene-specific values above the 95th (and lower than the 5th) quantile of expression were imputed to the 95th (5th) quantile to minimize the effect of outliers. Oscope was applied on the data with permuted sample order (**Supplementary Table 5**). After applying the paired-sine model to all genes, we used the top 5% as input for the *K*-medoids algorithm. Using the 151 genes in the top cluster (**Supplementary Table 6**), the ENI algorithm was applied with $m = 4$, and the degree of freedom of SPR was set to 3. To obtain the optimal order, we applied the 2-opt algorithm with 20,000 iterations (**Supplementary Table 7**). 874 genes were defined as periodic by the autoregression model in Whitfield *et al.*[8]. We used these 874 genes as a validation set in our evaluation.

H1 hESC cell culture. Undifferentiated H1 human embryonic stem cells (hESCs) were cultured in E8 medium[16] on Matrigel-coated tissue culture plates with daily media feeding at 37 °C with 5% (vol/vol) CO_2. Cells were split every 3–4 d with 0.5 mM EDTA in 1× PBS for standard maintenance. Immediately before single-cell suspensions for each experiment were prepared, hESCs were individualized by Accutase (Life Technologies), washed once with E8 medium and resuspended at densities of 5.0×10^5 to 8.0×10^5 cells/mL in E8 medium for cell capture. The H1 hESC line is registered in the NIH Human Embryonic Stem Cell Registry with the approval number NIHhESC-10-0043. Details of the H1 cells can be found online (http://grants.nih.gov/stem_cells/registry/current.htm?id=29). All cell cultures performed in our laboratory have routinely tested negative for mycoplasma contamination and have been authenticated by cytogenetic tests.

H1 hESC single-cell capture and single-cell cDNA library preparation. Single-cell loading, capture and library preparations were performed following the Fluidigm user manual "Using the C1 Single-Cell Auto Prep System to Generate mRNA from Single Cells and Libraries for Sequencing." Briefly, 5,000–8,000 cells were loaded onto a medium-sized (10- to 17-μm) C1 Single-Cell Auto Prep IFC (Fluidigm), and the cell-loading script was used according to the manufacturer's instructions. The capture efficiency was inspected using EVOS FL Auto Cell Imaging system (Life Technologies) to perform an automated area scanning of the 96 capture sites on the IFC. Empty capture sites or sites having more than one cell captured were first noted, and those samples were later excluded from further library processing for RNA-seq. Immediately after capture and imaging, reverse transcription and cDNA amplification were performed in the C1 system using the SMARTer PCR cDNA Synthesis kit (Clontech) and the Advantage 2 PCR kit (Clontech) according to the instructions in the Fluidigm user manual. Full-length, single-cell cDNA libraries were harvested the next day from the C1 chip and diluted to a range of 0.1–0.3 ng/μL. Diluted single-cell cDNA libraries were fragmented and amplified using the Nextera XT DNA Sample Preparation Kit and the Nextera XT DNA Sample Preparation Index Kit (Illumina). Libraries were multiplexed at 24 libraries per lane, and single-end reads of 67-bp were sequenced on an Illumina HiSeq 2500 system.

H1 hESC: read mapping and quality control. Reads were mapped against the hg19 RefSeq reference via Bowtie 0.12.8 (ref. 17) allowing up to two mismatches and up to 20 multiple hits. The expected counts and TPMs were estimated via RSEM 1.2.3 (ref. 18). Cells having fewer than 5,000 genes with TPM >1 were removed in quality control. 62, 78 and 73 cells passed the quality control in three replicate hESC experiments for a total of 213 H1 hESCs.

H1 hESC: statistical analysis. Expression within each cell was normalized following median normalization[19] implemented in EBSeq 1.5.4 (ref. 20). Gene means and variances were also estimated using EBSeq after adjusting for library sizes. High-mean and high-variance genes were selected before applying Oscope. Specifically, we took genes with mean expected count greater than 100 as genes with high mean. To define high-variance genes, we fit a linear model on log(variance) ~ log(mean) + *c*. Genes with variance above the fitted line were defined as high-variance genes. Genes with mappability scores[18] less than 0.8 were further eliminated. Applying these steps to the 213 H1 hESCs gave 2,376 genes to which Oscope was applied (**Supplementary Table 8**). Gene-specific values above the 95th (and below the 5th) quantile of expression were imputed to the 95th (5th) quantile to minimize the effect of outliers. After applying the paired-sine model, we used the top 5% of genes as input for the *K*-medoids algorithm. Using the 29 genes in the cell-cycle cluster, the ENI module was applied with $m = 4$, and the degree of freedom of SPR was set to 3. To obtain the optimal order, we applied the 2-opt algorithm with 20,000 iterations (**Supplementary Table 9**).

H1-FUCCI hESC cell line. Fluorescent ubiquitination-based cell-cycle indicator (FUCCI) H1 hESCs were generated by *piggyBac* insertion of a cassette encoding an *eef1a* promoter–driven *mCherryCDT1-IRES-EgfpGMNN* double transgene (custom ordered from GenScript). Individual clones were isolated by sorting double-positive single cells by fluorescence activated cell sorting (FACS) and maintained as described above. The H1-FUCCI cell line provides a two-color fluorescence labeling system allowing single-cell suspensions from G1, S or G2/M cell-cycle phases to be isolated by FACS. After this, single-cell suspensions were loaded onto the Single-Cell Auto Prep IFC using a medium-sized (10- to 17-μm) chip. FACS was performed on the FACSAria IIIu instrument and using FACSDiva software version 6.1.3 (both from Becton Dickinson). Unlabeled H1 cells or cells stained with single fluorochromes served as controls for fluorescence gating. Libraries and sequencing reads were processed in the same manner as described above.

H1-FUCCI: read mapping, quality control and statistical analysis. Reads were processed in the same way as in the H1 hESC data. A total of 91, 80, and 76 cells in G1, S and G2/M, respectively, passed our quality-control criteria as defined in the H1 hESC read-mapping and quality-control section. Statistical analysis on H1 and H1-FUCCI combined data was carried out as described in H1 hESC statistical analysis. The phase boundaries (**Fig. 2d**) are defined as the boundaries that give the smallest misclassification rate between three cell-cycle phases based on the reconstructed order (**Supplementary Table 10**).

Statistical model to identify genes with ordering effects. We used an ANOVA model to identify genes with potential ordering effects. Within each H1 hESC experiment, we grouped cells into eight groups defined by capture site. Recall that capture sites are labeled as A01, …, A12, B01, …, B12, …, H01, …, H12 to match their corresponding position in the output wells (**Fig. 3a** and **Supplementary Table 11**). We grouped cells from sites with the same starting letters. For each gene, we applied an ANOVA model on the combined data set from all three H1 hESC experiments. The model tests for differences in mean expression across the eight cell groups. A total of 403 genes were identified (P value < 0.005) using this ANOVA approach.

Single-cell real-time quantitative PCR (qPCR). Single-cell cDNA harvested from the Fluidigm C1 IFC was transferred to a 96-well plate and subsequently quantified and diluted according to the Fluidigm user manual. Two microliters of the diluted single-cell cDNA were subsequently used in replicated qPCR reactions with individual 1× TaqMan Gene Expression assays and 1× TaqMan Universal PCR Master Mix II (Life Technologies) in a total volume of 10.0 μL. qPCR was performed using ViiA 7 System, and data analysis was performed using ExpressionSuite (all from Life Technologies). TaqMan Gene Expression assays (Life Technologies) were used for two genes: *PFN1* (Hs00748915_s1) and *MIF* (Hs00236988_g1), with *GAPDH* (Hs02758991_g1) as an internal control. Although the TaqMan Gene Expression assay is compliant with the MIQE guidelines for publications, the actual sequences of the primers and probes are not released for each assay. The amplicon context sequence for each assay can be identified as follows: *PFN1*: 223 bp, 5′- ccaccttcggcgttcccagtactgacctcgtctgtcc cttccccttcaccgctccccacagctttgcaccccttttcctccccatacacacacaaaccatttt atttttgggccattaccccataccccttattgctgccaaaaccacatgggctgggggccagg gctggatggacagacacctccccctacccatatccctcccgtgtgtggttggaaaact-3′. *MIF*: 83 bp, 5′-ctgtgcggcctgctggccgagcgcctgcgcatcagcccggacagg gtctacatcaactattacgacatgaacgcggccaatgt-3′. *GAPDH*: 110 bp, 5′-cc

doi:10.1038/nmeth.3549

ctggccaaggtcatccatgacaactttggtatcgtggaaggactcatgaccacagtcca
tgccatcactgccacccagaagactgtggatggcccctccgggaaactg-3′.

Code availability. The R package R/Oscope is available at https://www.biostat.wisc.edu/~kendzior/OSCOPE/.

16. Chen, G. *et al. Nat. Methods* **8**, 424–429 (2011).
17. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. *Genome Biol.* **10**, R25 (2009).
18. Li, B. & Dewey, C.N. *BMC Bioinformatics* **12**, 323 (2011).
19. Anders, S. & Huber, W. *Genome Biol.* **11**, R106 (2010).
20. Leng, N. *et al. Bioinformatics* **29**, 1035–1043 (2013).