Uncovering cellular processes in HeLa cells using unsupervised machine learning

Tom Maimon[1], Yaron Trink[1], Jacob Goldberger[1], and Tomer Kalisky [1,2]

[1]*Faculty of Engineering and Bar-Ilan Institute of Nanotechnology and Advanced Materials (BINA), Bar-Ilan University, Ramat Gan, Israel 5290002*

[2] **Correspondence to:**

Tomer Kalisky

Faculty of engineering, Bar-Ilan University, Ramat Gan, Israel 52900

Office: +972-3-738-4656

Email: tomer.kalisky@biu.ac.il

ABSTRACT

Gene expression measurements taken over multiple time points are useful for describing dynamic biological phenomena such as tissue development, regeneration, and cancer. However, since these phenomena involve multiple processes occurring in parallel, for example differentiation and proliferation, it is difficult to discern their respective contributions to a measured gene expression profile at any given point in time. Here, we demonstrate the use of un-supervised machine learning techniques to identify and "de-convolve" processes occurring in parallel in a simple model system. We first downloaded a published dataset of RNAseq measurements from synchronized HeLa cells that were sampled at 14 consecutive time points. We then used Fourier analysis and Topic modeling to identify two concurrent processes: a periodic process, corresponding to cell division, and a transient process related to HeLa cell identity (e.g. cervical cancer), that is presumably required for recovery from cell cycle arrest. This study demonstrates the use of un-supervised machine learning techniques to identify hidden states and processes in the cell.

KEYWORDS

Cell cycle, Fourier transform, Topic modeling, machine learning, HeLa cells

INTRODUCTION

Many biological phenomena such as development, regeneration, and cancer, involve thousands of cells dynamically changing their transcriptional state over time, as multiple genes are activated and repressed through a complex network of interactions. One way to characterize cell state dynamics is to perform RNA sequencing at multiple time points, either at the “bulk” or single-cell level, and obtain a series of gene expression profiles that create continuous trajectories in a latent gene expression “space”. However, cell states are often a result of multiple processes occurring in parallel, for example, differentiation and proliferation, and the relative contributions of each process to the measured gene expression profiles are often difficult to ”deconvolve” from each other.

Here, we demonstrate the use of an unsupervised machine learning technique for identifying and “de-convolving” processes in a simple model system. We first downloaded a dataset of bulk RNAseq measurements performed by Dominguez *et al.* on synchronized HeLa cells at 14 consecutive time points, which corresponds to approximately two cell cycles [1,2]. Then, we performed Fourier analysis and found that the majority of genes have either a periodic or a transient behavior over time. Finally, we used Topic modeling - an unsupervised machine learning technique - to show that the series of gene expression profiles can be represented as a mixture of three hidden topics, two of which are periodic, corresponding to the G1-S and G2-M phases of the cell cycle, and one with a transient behavior that is related to HeLa cell identity (e.g. cervical cancer), presumably required for recovering for cell cycle arrest. We believe that this study demonstrates the use of machine learning algorithms to deconvolve complex biological processed and identify their hidden components.

MATERIALS AND METHODS

Data:

The RNAseq dataset generated by Dominguez *et al.* [1,2] was downloaded from GEO (accession number GSE81485) using SRA tools.

Preprocessing:

STAR [3] was used to align the reads to the human reference genome (hg38) and to obtain the raw counts matrix. DESeq2 [4] was used to obtain the normalized counts matrix. Then, genes whose count sum across all 14 samples was below 10 were removed.

RNA velocity:

Velocyto [5] was used to obtain the read counts matrices of the spliced and unspliced mRNA. DESeq2 normalization was performed for the two matrices combined.

Fourier Transform:

For each gene, Fourier analysis was performed using the R function periodogram() from the "TSA" (Time Series Analysis) R package. This function returns a vector of frequencies and a vector of scores (spectral densities) for each frequency. For each gene we found the "dominant frequency", that is, the frequency with the highest score, and calculated a "dominant frequency score" which measures the degree to which the score of this frequency is higher than the scores of the other frequencies. Specifically, if for a particular gene the highest score is S1 and its matching frequency is F1, the second highest score is S2 and its frequency is F2, the third ranking score is S3 and its frequency is F3 etc., then the "dominant frequency" for that gene is F1, and the "dominant frequency score" is: $S1/(S2+S3)$.

Topic modeling:

Topic modelling was performed on the normalized counts matrix using the R functions fit_topic_model(), diff_count_analysis(), and structure_plot() from the "fastTopics" R package [6] with k=3 topics, which we labeled as "k1", "k2", and "k3". Using these

88 functions, we calculated the three probability distributions $p(gene|k1)$, $p(gene|k2)$,
89 and $p(gene|k3)$ for every gene in each one of the three topics, as well as the
90 probabilities $p(k1|sample)$, $p(k2|sample)$, and $p(k3|sample)$ for each topic in every
91 one of the fourteen samples. We also tested higher numbers of topics but the results
92 did not change significantly.
93
94 One way to visualize the association between a specific gene and each of the three
95 topics is to calculate the "posterior probabilities" $p(k1|gene)$, $p(k2|gene)$, and
96 $p(k3|gene)$ and compare them. In order to do this, we used Bayes' theorem:

97 $$p(topic|gene) = \frac{p(gene|topic) \cdot p(topic)}{\sum_{topic} p(gene|topic) \cdot p(topic)}$$

98 where, for simplicity, we set the prior $p(topic) = 1/3$ for each topic.
99
100 **RESULTS**
101 **RNA velocity analysis of periodically expressed genes reveals a time lag between**
102 **spliced and un-spliced mRNA:**
103 We downloaded a published dataset of RNAseq measurements that were performed by
104 Dominguez *et al.* [1,2]. In that study, HeLa cells were first synchronized by a double
105 thymidine block and then RNA was collected and sequenced at 14 consecutive time
106 points, which corresponds to approximately two cell cycles (Figure 1A, Table S1). The
107 authors also identified a set of 67 "core" cell cycle related genes with periodic behavior,
108 and categorized them as "G1-S related" or "G2-M related" according to the stage at
109 which their expression is maximal (Table S1).
110
111 We started by manually inspecting the expression levels of spliced and un-spliced mRNA
112 of "core" cell cycle genes such as CCNE2 and UNG (G1-S related) and MKI67 and TOP2A
113 (G2-M related, Figure 1B) [5]. We observed that the yet-unspliced mRNA typically
114 precedes the spliced mRNA by 2-3 hours, as expected. We confirmed this by performing
115 a cross-correlation between the spliced and un-spliced mRNA levels averaged over the
116 67 "core" cell cycle genes (Figure 1C).

Fourier analysis identifies sets of genes with periodic and transient behaviors over time:

In order to test the periodicity of every gene in the transcriptome, we performed Fourier analysis. For each gene we plotted its "dominant frequency", that is, the frequency with the highest spectral density, versus its "dominant frequency score", which measures the degree to which the score of this frequency is higher than the scores of the other frequencies (see Methods and Figure 2A). We found that the first or second dominant frequencies - that correspond to either one or two whole cycles – contain genes with higher scores than the other dominant frequencies, whose genes have scores that are similar to those of randomized datasets (Figure S1). This indicates that most of the information in our dataset is contained in the expression levels of the genes having one of the first two dominant frequencies.

We next selected genes from the second dominant frequency (with scores > 3) and performed PCA. We found that the samples form a circular trajectory in latent space that corresponds to almost two complete cell cycles (Figure 2B). Moreover, separating the spliced and yet-unspliced mRNA expression profiles results in two circular trajectories in latent space with a phase difference between them (Figure S2). When we selected the genes from the first and second dominant frequencies we found that the samples form a trajectory in latent space with two patterns of behavior: a transient pattern along PC1 (Figure 2C) and an oscillatory pattern along PC2 and PC3 (Figure 2D-E).

Topic modeling shows that the time-dependent series of gene expression profiles can be represented as a mixture of three hidden topics, two of which are periodic, corresponding to the G1-S and G2-M phases of the cell cycle, and one with a transient behavior that is related to HeLa cell identity (i.e. cervical cancer):

145 “Topic models” are used in computer science to model documents that contain words
146 from different “topics” [7,8]. Given a set of documents and an assumed number of
147 topics k, the model assumes that each document can be described as a mixture of these
148 k latent topics with different proportions. By fitting a topic model to that set of
149 documents, it is possible to discover both the k latent topics and their proportions in
150 each individual document. In our case we assume that each cell state (=gene expression
151 profile) is a weighted sum of k hidden “processes” (that is, latent components), and that
152 these weights vary at each time point. Therefore, if each gene expression profile
153 corresponds to a document and each process corresponds to a topic, fitting a topic
154 model to our data can help us discover both the k latent processes and their proportions
155 at each individual time point.
156
157 We therefore fitted a topic model with K=3 topics to the 14 gene expression profiles and
158 found three processes with distinct behaviors, which we labeled as “k1”, “k2” and “k3”
159 (Figure 3A and Table S2). We observed that topics k1 and k3 have an oscillatory
160 behavior and are periodically expressed at specific phases of the cell cycle throughout
161 the entire duration of the experiment (approximately two cell cycles) (Figure 3B),
162 whereas topic k2 is transiently over-expressed mainly during the first cycle following
163 release from the thymidine block.
164
165 Next, we used ToppGene [9] to perform Gene Ontology (GO) enrichment analysis on
166 genes that are significantly over-expressed in each one of the three topics in order to
167 identify them (Figure 4A-B, Tables S3-S5). We found that topic k1 over-expresses genes
168 that are known to be associated with the G1-S phase of the cells cycle (e.g. DNA
169 replication), topic k3 over-expresses genes that are known to be associated with the G2-
170 M phase of the cells cycle (e.g. mitosis), and topic k2 over-expresses genes related to
171 HeLa cell identity (e.g. cervical carcinoma).
172

173 We confirmed this by calculating the posterior probabilities p(k1|gene), p(k2|gene), and
174 p(k3|gene) for selected genes with specific GO terms (Figure 4C). These probabilities
175 represent the association between these genes and each of the three topics. Indeed, we
176 observed that within the 67 "core" cell cycle genes, the G1-S related genes were
177 associated with topic k1 and the G2-M related genes were associated with topic k3.
178 Similarly, genes related to HeLa cell identity and Cervical Carcinoma were more
179 associated with topic k2.
180
181 **DISCUSSION**
182 To summarize, we applied Fourier analysis and Topic modeling to a simple *in-vitro*
183 model system in order to identify its latent components. We identified two periodic
184 processes corresponding to the G1-S and G2-M phases of the cell cycle and one
185 transient process related to HeLa cell identity and cervical cancer. Our results suggest
186 that, following release from cell cycle arrest induced by the double thymidine block, the
187 cells enter a transient state lasting approximately twelve hours (samples D1-D7) in
188 which they activate molecular mechanisms in order to exit cell cycle arrest and resume
189 cycling. Since each tissue regulates cell division in a unique way, it is logical that many of
190 these mechanisms are specific to the tissue from which HeLa cells originated from (i.e.
191 cervical cancer).
192
193 We note that for some of the 67 "core" cell cycle genes we observed that the spliced
194 and unspliced mRNA levels arrive at their first peak (maximum or minimum) at the same
195 time (Figure 1B), which indicates that initially, during the few hours after release from
196 the thymidine block, the spliced and unspliced transcripts are synchronized in these
197 genes, and only after a few hours the typical time lag is acquired. This also indicates the
198 existence of a transient cell state in which the transcriptional timing mechanisms
199 operate in a way that is different from the normal cycling state. One possible
200 explanation is that during the first few hours after release from cell cycle arrest, the
201 primary mRNA transcribed from these genes is spliced, processed, and translated with

minimal delay in order to rebuild the necessary mechanisms for cell cycle resumption as soon as possible. Further experimental studies are required to fully resolve and validate these mechanisms.

ACKNOWLEDGEMENTS

We wish to thank Shahar Alon and all the member of our lab for useful comments and suggestions.

DECLARATION OF INTEREST STATEMENT

The authors have declared that no competing interests exist.

AUTHOR CONTRIBUTIONS

Study initiation and conception – T.M. and T.K.; Data analysis - T.M. and T.K.; Topic modeling – T.M, J.G., and T.K.; Other intellectual contribution – Y.T.; Manuscript writing – T.M. and T.K.

FUNDING

T.M., Y.T., and T.K. were supported by the Israel Science Foundation (ICORE no. 1902/12 and Grants no. 1634/13 and 2017/13), the Israel Cancer Association (Grant no. 20150911), the Israel Ministry of Health (Grant no. 3-10146), the EU-FP7 (Marie Curie International Reintegration Grant no. 618592), and the ICRF (Grant no. 19-101-PG). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

226 **REFERENCES**

227 1. Dominguez D, Tsai YH, Weatheritt R, Wang Y, Blencowe BJ, Wang Z. An extensive
228 program of periodic alternative splicing linked to cell cycle progression. Elife.
229 2016;5: 1–19. doi:10.7554/eLife.10288
230 2. Dominguez D, Tsai YH, Gomez N, Jha DK, Davis I, Wang Z. A high-resolution
231 transcriptome map of cell cycle reveals novel connections between periodic
232 genes and cancer. Cell Res. 2016;26: 946–962. doi:10.1038/cr.2016.84
233 3. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast
234 universal RNA-seq aligner. Bioinformatics. 2013;29: 15–21.
235 doi:10.1093/bioinformatics/bts635
236 4. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion
237 for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550. doi:10.1186/s13059-
238 014-0550-8
239 5. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA
240 velocity of single cells. Nature. 2018;560: 494–498. doi:10.1038/s41586-018-
241 0414-6
242 6. Carbonetto P, Luo K, Dey K, Hsiao J, Stephens M. fastTopics: fast algorithms for
243 fitting topic models and non-negative matrix factorizations to count data. R
244 package version 0.4-11. 2021. Available:
245 https://github.com/stephenslab/fastTopics
246 7. Dey KK, Hsiao CJ, Stephens M. Visualizing the structure of RNA-seq expression
247 data using grade of membership models. PLoS Genet. 2017;13: 1–23.
248 doi:10.1371/journal.pgen.1006599
249 8. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. J Mach Learn Res. 2003;3:
250 993–1022. Available: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
251 9. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment
252 analysis and candidate gene prioritization. Nucleic Acids Res. 2009.
253 doi:10.1093/nar/gkp427

254

APPENDICES

Supplementary information: Supplementary text and figures.

Table S1: The gene expression dataset from Dominguez *et al.*, including the sampling time points and the list of 67 "core" cell cycle genes.

Table S2: The three topics discovered and their proportions at each time point.

Table S3: GO enrichment analysis for genes that are over-expressed in topic k1.

Table S4: GO enrichment analysis for genes that are over-expressed in topic k2.

Table S5: GO enrichment analysis for genes that are over-expressed in topic k3.

Program: A compressed directory containing programs and datasets for data visualization.

FIGURES:

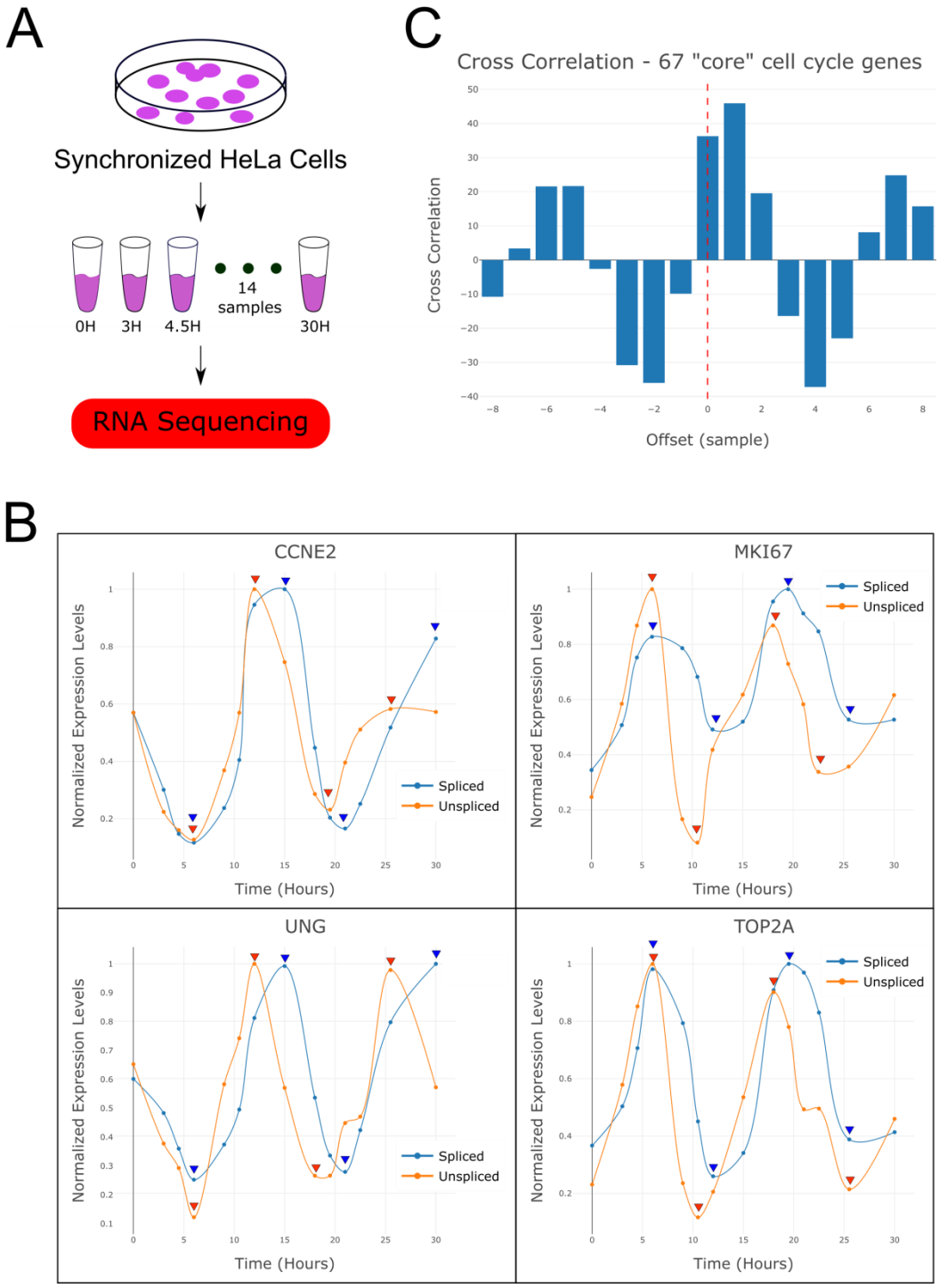


Figure 1: RNA velocity analysis of periodically expressed genes reveals a time lag between spliced and un-spliced mRNA. (A) Shown is a sketch of the experiment

271 performed by Dominguez *et al.* from which the RNAseq dataset was produced. HeLa
272 cells were synchronized, sampled at 14 consecutive time points, and their RNA was
273 sequenced. (B) Shown are expression levels over time of spliced and un-spliced mRNA
274 for selected genes that are known to be over-expressed at the G1-S phases (CCNE2,
275 UNG) or G2-M phases (MKI67, TOP2A) of the cell cycle. These indicate that the cells
276 completed two cell divisions during the experiment. It can be seen that the yet un-
277 spliced mRNA again precedes the spliced mRNA, as expected. Expression levels were
278 normalized to their maximum and spline curves were plotted through the data points to
279 assist visibility. (C) A cross-correlation plot between the spliced and un-spliced mRNA
280 levels averaged over the 67 "core" cell cycle genes, that is, periodically expressed genes
281 identified by Dominguez *et al.* to be related to the cell cycle in this dataset, shows a
282 significant time lag between the spliced and yet un-spliced mRNA.
283

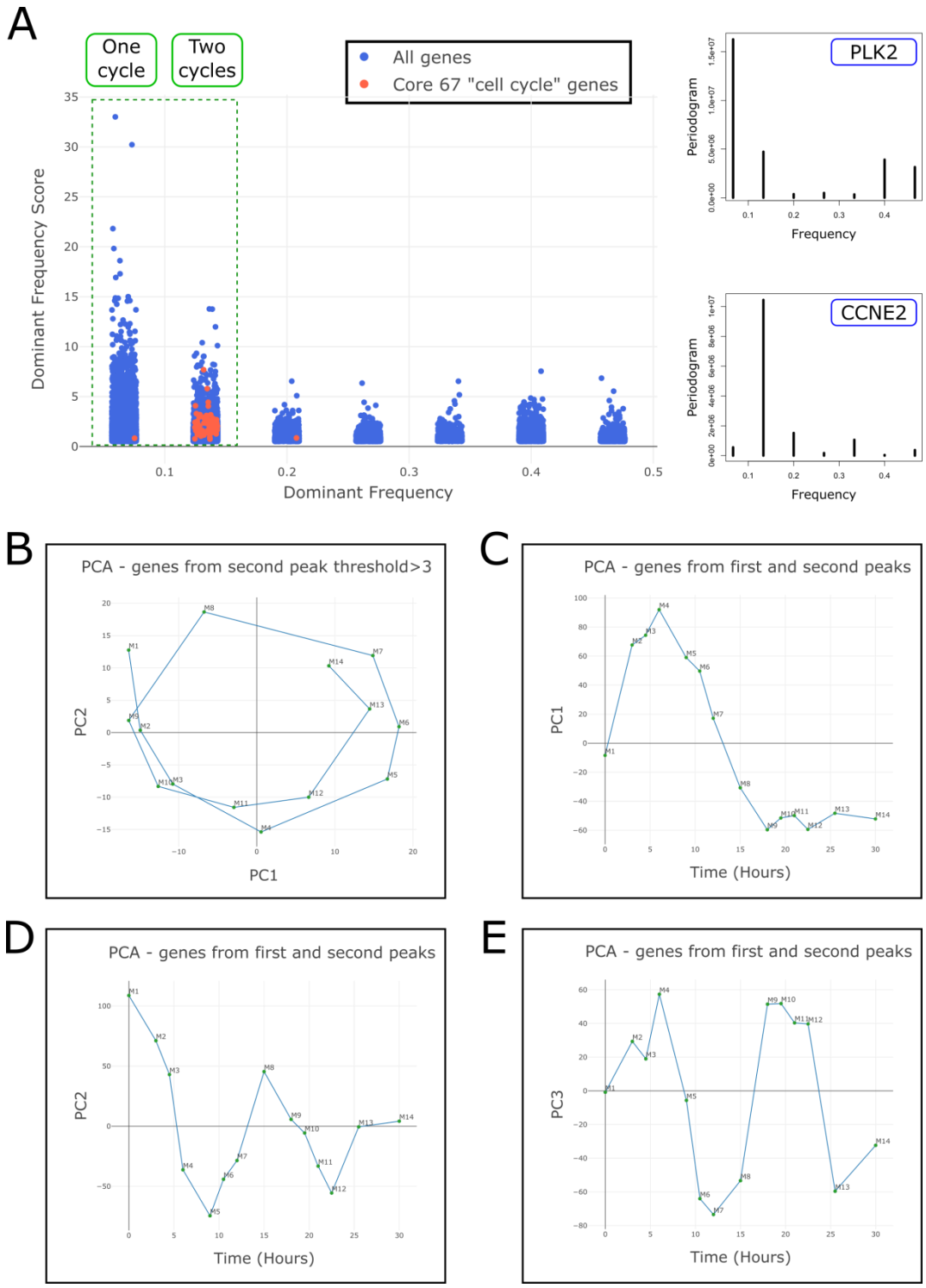


Figure 2: Fourier analysis identifies sets of genes with periodic and transient behaviors over time. (A) For each gene, a periodogram was calculated (e.g., PLK2 and CCNE2, right panels) and its "dominant frequency" vs. "dominant frequency score" were plotted (left panel, see Methods). It can be seen that almost all of the "core" cell cycle genes (65 out

of 67 red dots) have a dominant frequency corresponding to two whole cycles. Likewise, the first and second dominant frequencies contain genes with higher scores than the other frequencies, whose genes have scores that are similar to those of randomized datasets (Figure S1), indicating that the genes within the first and second dominant frequencies contain most of the periodic information in our dataset. (B) When we select genes from the second dominant frequency with scores > 3 and perform PCA, the samples form a circular trajectory in latent space that represents almost two complete cell cycles. (C-E) When we select genes from the first and second dominant frequencies and perform PCA, the samples form a trajectory in latent space with two patterns of behavior: a transient pattern (along PC1 axis) and a periodic pattern (in the PC2 vs. PC3 plane).

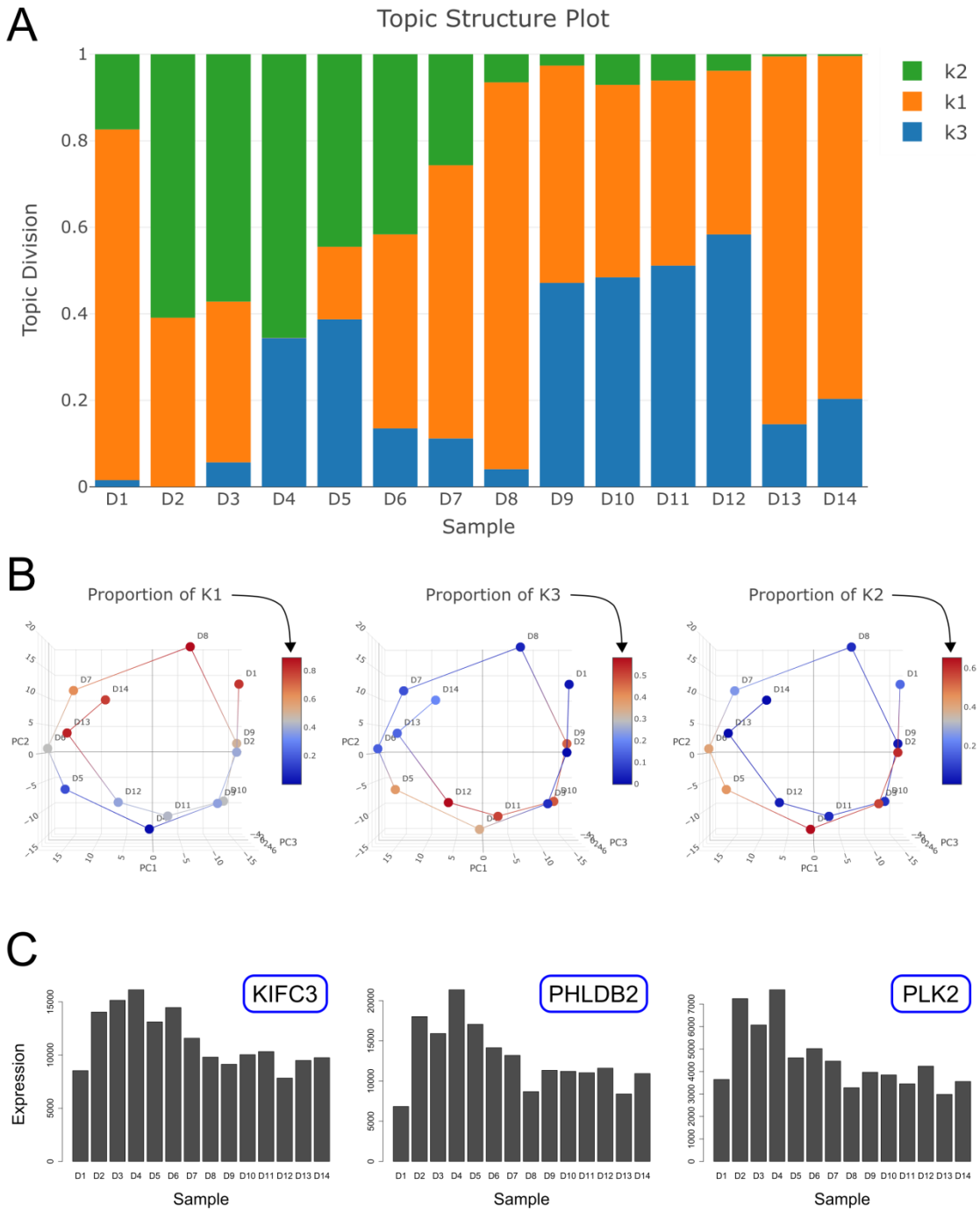
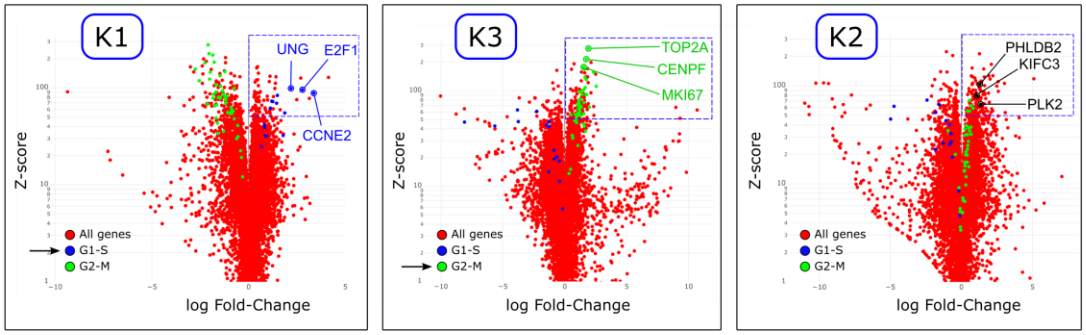


Figure 3: Topic modeling shows that the time-dependent series of gene expression profiles can be represented as a mixture of three hidden topics, two of which are periodic and one with a transient behavior. (A) Shown is a structure plot of the proportions of each topic within each sample. Topics k1 and k3 are periodically expressed during both cycles (samples D1-D14), whereas k2 is transiently over-

308 expressed during the first cycle (samples D1-D7). (B) Shown is a visualization of the
309 proportions of each topic along the circular trajectory in latent space representing the
310 two cell cycles. It can be seen that topics k1 and k3 are over-expressed in specific phases
311 during both cycles, whereas topic k2 is over-expressed during the first cycle only. Note
312 that topic k1 is expressed mainly in the upper half of the circular trajectory, whereas
313 topic k3 is expressed mainly in its lower half. (C) Shown are the expression levels of
314 three representative genes that appear at high probability in topic k2. These genes are
315 transiently over-expressed in the first cycle.
316

A



B

Name	Source	ID	Bonferroni	Hit count	Total list
Genes whose expression peaked at 480 min after stimulation of HeLa cells with EGF [GeneID=1950].	MSigDB C2: CGP Curated Gene Sets (v7.3)	M8729	6.13E-26	26	164
Genes modulated in HeLa cells (cervical carcinoma) by TNF [GeneID=7124] via NFKB pathway.	MSigDB C2: CGP Curated Gene Sets (v7.3)	M14435	1.80E-10	9	26
TCGA-Cervix-Primary_Tumor-Cervical_Carcinoma-Cervical_Squamous_Cell_Carcinoma-6|TCGA-Cervix / Sample_Type by Project: Shred V9	TCGA OncoMap	45a8e069369fdc7358135bf2229defbee1ac2600	7.44E-06	9	111
TCGA-Cervix-Primary_Tumor-Cervical_Carcinoma-Non-keratinizing_Cervical_Squamous_Cell_Carcinoma-6|TCGA-Cervix / Sample_Type by Project: Shred V9	TCGA OncoMap	10955b54e4a87d035093a2756291ef1d6a14ab63	1.30E-03	8	143

C

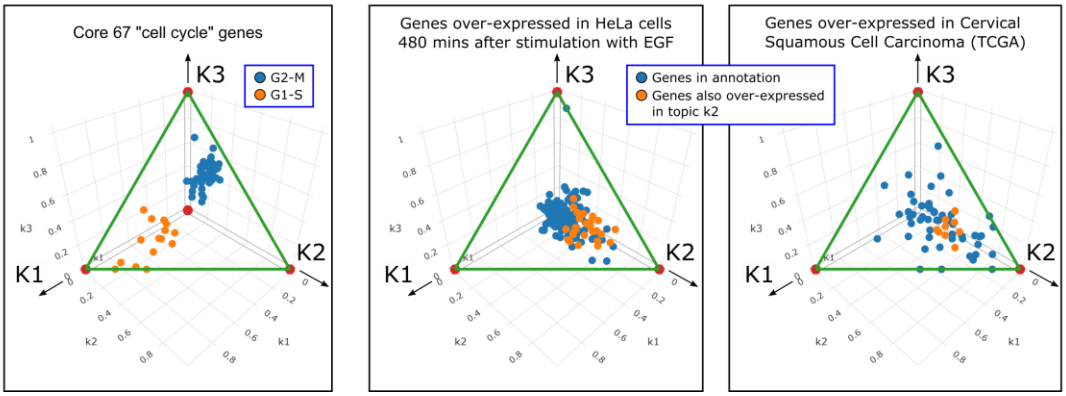


Figure 4: Gene Ontology (GO) enrichment analysis shows that the three hidden topics are related to the G1-S/G2-M phases of the cell cycle and to HeLa cell identity (i.e. cervical cancer). (A) Volcano plots showing the over-expression of each gene in each of the three topics. It can be seen that genes known to be over-expressed in the G1-S phases of the cell cycle are over-expressed in topic k1 (left panel), genes known to be over-expressed in the G2-M phases of the cell cycle are over-expressed in topic k3 (middle panel), and genes that are transiently over-expressed in the first cycle are over-expressed in topic k2 (right panel). (B) GO enrichment analysis of genes over-expressed

327 in topic k2 (Z-score>50) shows that topic k2 is related to HeLa cell identity and cervical
328 cancer (see also Tables S3-S5). (C) Shown are the posterior probabilities $p(topic|gene)$
329 for selected gene sets with specific GO terms. Each point represents a gene, and its
330 components along the k1, k2, and k3 axes are the posterior probabilities $p(k1|gene)$,
331 $p(k2|gene)$, and $p(k3|gene)$, which represent the association between that gene and
332 each of the three topics. It can be seen that topic k1 is highly associated with genes that
333 are known to be over-expressed in the G1-S phases of the cell cycle (left panel), whereas
334 topic k3 is highly associated with genes that are known to be over-expressed in the G2-
335 M phases of the cell cycle. Topic k2 is associated with genes that are related to HeLa cell
336 identity and Cervical Carcinoma (middle and right panels). Note that all points (=genes)
337 are located on the triangle-shaped two dimensional simplex since the sum of the
338 posterior probabilities for each gene equals 1.
339