Unsupervised machine learning reveals temporal components of gene expression in HeLa cells following release from cell cycle arrest

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 of tissues, tumors, or cell lines taken over multiple time points are valuable for describing dynamic biological phenomena such as the response to growth factors. However, such phenomena typically involve multiple biological processes occurring in parallel, making it difficult to identify and discern their respective contributions at any time point. Here, we demonstrate the use of unsupervised machine learning to deconvolve a series of time-dependent gene expression measurements into its underlying temporal components. We first downloaded publicly available RNAseq data obtained from synchronized HeLa cells at consecutive time points following release from cell cycle arrest. Then, we used Fourier analysis and Topic modeling to reveal three underlying components and their relative contributions at each time point. We identified two temporal components with oscillatory behavior, corresponding to the G1-S and G2-M phases of the cell cycle, and a third component with a transient expression pattern, associated with the immediate-early response gene program, regulation of cell proliferation, and cervical cancer. This study demonstrates the use of unsupervised machine learning to identify hidden temporal components in biological systems, with potential applications to early detection and monitoring of diseases and recovery processes.

KEYWORDS

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

INTRODUCTION

Complex biological systems such as tissues and tumors contain millions of cells. These cells dynamically change their transcriptional state over time as multiple gene programs - coordinated sets of genes that work together to perform specific biological tasks - are activated and repressed through a complex network of interactions. Dynamic changes in transcriptional cell states underlie fundamental biological processes, such as embryo development, tissue regeneration, and the onset and progression of cancer. A common approach to characterize cell state dynamics is to perform RNA sequencing at multiple time points, either at the 'bulk' or single-cell level, resulting in a series of gene expression profiles that represent the underlying cell states and that typically form continuous trajectories in latent space. However, the resulting expression profiles are typically a combination of multiple time-dependent components, where each component is associated with distinct biological processes and gene programs. A major challenge is to “deconvolve” these temporal components. This involves characterizing each component, identifying its associated gene programs, and inferring its relative contribution to the overall gene expression profiles at each time point.

Here, we demonstrate the use of unsupervised machine learning to reveal the temporal components in a simple model biological system. We first downloaded a dataset of bulk RNAseq measurements performed by Dominguez *et al.* on synchronized HeLa cells at 14 consecutive time points - approximately two cell cycles - following release from cell cycle arrest (Dominguez, Tsai, Gomez, et al. 2016; Dominguez, Tsai, Weatheritt, et al. 2016). Then, we performed Fourier analysis and found that most genes have a periodicity of either one or two cycles over time. Next, we used topic modeling, an unsupervised machine learning technique, and found that the series of gene expression profiles can be represented as a mixture of three components, two of which are periodic over time and correspond to the G1-S and G2-M phases of the cell cycle, and a third topic with a transient temporal behavior, that is associated with the immediate-early response gene program, regulation of cell proliferation, and cervical cancer. This study

demonstrates the potential of machine learning algorithms to deconvolve hidden temporal components in complex biological systems, with potential applications for early disease detection, as well as monitoring disease progression and recovery processes.

MATERIALS AND METHODS

Datasets and Preprocessing:

The RNAseq dataset generated by Dominguez *et al.* (Dominguez, Tsai, Gomez, et al. 2016; Dominguez, Tsai, Weatheritt, et al. 2016) was downloaded from GEO (accession number GSE81485) using SRA tools. STAR (Dobin et al. 2013) was used to align the reads to the human reference genome (hg38) and to obtain the raw counts matrix. DESeq2 (Love, Huber, and Anders 2014) was used to obtain the normalized counts matrix. Genes whose sum of counts across all 14 samples was below 10 were removed.

RNA velocity:

Velocyto (La Manno et al. 2018) was used to obtain the read counts of the spliced and unspliced mRNA for each sample. DESeq2 normalization was performed for the spliced and unspliced counts matrixes combined.

Fourier Transform:

For each gene, a 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: $S1/(S2+S3)$. The significance of the dominant frequency scores was tested by comparing them to scores derived from randomized datasets, which were obtained by randomly shuffling the order of counts for each gene.

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 (Carbonetto et al. 2021). We fitted a topic model with k=3 topics that we labeled as "k1", "k2", and "k3". Using these functions, we calculated the three probability distributions $p(gene|k1)$, $p(gene|k2)$, and $p(gene|k3)$ for every gene in each one of the three topics, as well as the probabilities $p(k1|sample)$, $p(k2|sample)$, and $p(k3|sample)$ for each topic in every one of the fourteen samples. We also tested higher numbers of topics but the results did not change significantly.

One way to visualize the association between a specific gene and each of the three topics is to calculate the "posterior probabilities" $p(k1|gene)$, $p(k2|gene)$, and $p(k3|gene)$ and compare them. In order to do this, we used Bayes' theorem:

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

where, for simplicity, we set the prior $p(topic) = 1/3$ for each topic.

Gene Ontology (GO) enrichment analysis

GO enrichment analysis was performed with Toppgene (Chen et al. 2007) (https://toppgene.cchmc.org/). Gene set enrichment analysis was performed with GSEA (Subramanian et al. 2005) (see Supplementary data for run details).

RESULTS

RNA velocity analysis of periodically expressed genes reveals a time lag between spliced and un-spliced mRNA:

120 We downloaded a published dataset of RNAseq measurements that were performed by
121 Dominguez *et al.* (Dominguez, Tsai, Gomez, et al. 2016; Dominguez, Tsai, Weatheritt, et
122 al. 2016). In that study, HeLa cells were first synchronized by a double thymidine block.
123 Then, following release from cell cycle arrest, RNA was collected and sequenced at 14
124 consecutive time points which corresponds to approximately two cell cycles (Figure 1A,
125 Table S1). The authors also identified a set of 67 "core" cell cycle related genes with
126 periodic behavior, and categorized them as "G1-S related" or "G2-M related" according
127 to the stage at which their expression is maximal (Table S1).
128
129 We started by manually inspecting the expression levels of spliced and un-spliced mRNA
130 of "core" cell cycle genes such as CCNE2 and UNG (G1-S related) and MKI67 and TOP2A
131 (G2-M related, Figure 1B) (La Manno et al. 2018). As expected, we observed that the
132 yet-unspliced mRNA typically precedes the spliced mRNA by a single sampling interval,
133 which corresponds to a time interval of approximately 1.5-3 hours. We confirmed this
134 by performing a cross-correlation between the spliced and un-spliced mRNA levels
135 averaged over the 67 "core" cell cycle genes (Figure 1C).
136
137 **Fourier analysis identifies sets of genes with potentially transient and oscillatory**
138 **behaviors over time:**
139 To test the periodicity of every gene in the transcriptome, we performed Fourier
140 analysis. For each gene we plotted its "dominant frequency", that is, the frequency with
141 the highest spectral density, versus its "dominant frequency score", which measures the
142 degree to which the score of this frequency is higher than the scores of the other
143 frequencies (see Methods and Figure 2A). We found that the first and second dominant
144 frequencies - that correspond to either one or two whole cycles – contain genes with
145 scores that are higher than those within other dominant frequencies (Figure 2A).
146 Moreover, we observed that genes with dominant frequencies corresponding to three
147 cycles and higher have scores comparable to those derived from randomized datasets,
148 which were generated by randomly shuffling the order of counts for each gene (Figure

149 S1). This suggests that most of the information in our dataset is contained in the
150 expression levels of genes with a periodicity of either one or two cycles, indicating
151 potential transient or oscillatory behavior, respectively.
152
153 We next selected genes from the second dominant frequency (with scores > 3) and
154 performed PCA. We found that the samples form a circular trajectory in latent space
155 that corresponds to almost two complete cell cycles (Figure 2B). Moreover, separating
156 the spliced and yet-unspliced mRNA expression profiles results in two circular
157 trajectories in latent space with a phase difference between them (Figure S2). When we
158 selected the genes from the first and second dominant frequencies we found that the
159 samples form a trajectory in latent space with two patterns of behavior: a transient
160 pattern along PC1 (Figure 2C) and an oscillatory pattern along PC2 and PC3 (Figures 2D-
161 E).
162
163 **Topic modeling reveals three temporal components, two of which are periodic,**
164 **corresponding to the G1-S and G2-M phases of the cell cycle, and a third, transient**
165 **component, related to immediate-early response, regulation of cell proliferation, and**
166 **cervical cancer:**
167 "Topic models" are used in computer science to describe collections of documents that
168 contain varying proportions of words from a predetermined number k of different
169 "topics" (Blei, Ng, and Jordan 2003; Dey, Hsiao, and Stephens 2017). By fitting a topic
170 model to a set of documents, it is possible to discover both the k latent topics (i.e., the
171 probability of occurrence of each word from the vocabulary within a topic) and their
172 proportions in each individual document. In our case we assume that each cell state
173 (gene expression profile) at a given time point is a weighted sum of k hidden
174 components whose proportions vary with time. We therefore treat each cell state at a
175 given time point as a document, each gene in the human genome as a word in the
176 vocabulary, the presence of a specific RNA transcript within a cell state as the
177 occurrence of a specific word in a document, the expression level (number of

178 transcripts) of a gene in a cell state as the number of occurrences of a word in a
179 document, and each latent component as a topic. Topic modeling can thus be used to
180 find the k components (i.e. the probability of observing an RNA transcript from each
181 gene within a component) and their proportions at each individual time point.
182
183 We fitted a topic model with k=3 topics to the 14 gene expression profiles and identified
184 three components/topics with distinct temporal patterns, which we labeled as "k1",
185 "k2" and "k3" (Figure 3A and Table S2). We found that topics k1 and k3 have an
186 oscillatory behavior over the course of both cycles following release from cell cycle
187 arrest and are periodically over-expressed at specific phases of the cell cycle (Figures 3A-
188 B). Topic k2 however is transiently over-expressed mainly during the first cycle following
189 release from cell cycle arrest. We confirmed that the temporal pattern of each topic is
190 also observed in selected genes associated with it. For example, the genes CCNE2 and
191 UNG whose expression is periodic are associated with topic k1 (Figures 1B and 4A),
192 whereas the genes MKI67 and TOP2A whose expression is also periodic, but with a
193 different phase, are associated with topic k3 (Figures 1B and 4A). Likewise, the genes
194 KIFC3, PHLDB2, and PLK2, which are associated with topic k2, are transiently over-
195 expressed following release from cell cycle arrest (Figures 3C and 4A).
196
197 To identify and characterize the topics, we performed Gene Ontology (GO) enrichment
198 analysis. For each topic, we selected genes that are significantly over-expressed and
199 used them as input to ToppGene (Chen et al. 2009) (Figure 4A, Tables S3-S5). We found
200 that topic k1 over-expresses genes related to the G1-S phase of the cells cycle (e.g. DNA
201 replication and G1 to S cell cycle control) and that topic k3 over-expresses genes that
202 are related to the G2-M phase of the cells cycle (e.g. mitosis and regulation of the G2/M
203 transition). In contrast, topic k2 was found to over-express genes associated with the
204 following gene programs: (i) immediate-early response (Herschman 1991; O'Donnell,
205 Odrowaz, and Sharrocks 2012; Tullai et al. 2007; Winkles 1997) (Table S6 and Figure S3);
206 (ii) regulation of cell proliferation (Table S4), and (iii) cervical carcinoma (Table S4).

207

208 We confirmed these observations by calculating the posterior probabilities p(k1|gene),
209 p(k2|gene), and p(k3|gene) for genes belonging to specific gene sets and GO terms
210 (Figure 4C). These probabilities represent the association between these genes and each
211 of the three topics. Indeed, we observed that within the 67 "core" cell cycle genes, the
212 G1-S related genes were associated with topic k1 and the G2-M related genes were
213 associated with topic k3. Likewise, genes related to immediate-early response and
214 cervical carcinoma were more strongly associated with topic k2.

215

216 **DISCUSSION**

217 In this study we applied Fourier analysis and Topic modeling to a simple *in-vitro* model
218 system in order to identify its latent components. We identified two components with a
219 periodic temporal behavior (topics k1 and k3), that correspond to the G1-S and G2-M
220 phases of the cell cycle, and a third component with a transient expression pattern
221 (topic k2) that is associated with the immediate-early response gene program,
222 regulation of cell proliferation, and cervical cancer.

223

224 Specific examples of genes that are transiently over-expressed following release from
225 cell cycle arrest, and that have a high probability for being expressed in topic k2, are
226 KIFC3, PHLDB2, and PLK2 (Figures 3C and 4A). These genes are known to be associated
227 with cervical carcinoma (Table S4). Moreover, PLK2, a member of the polo-like kinase
228 family, is also known to be associated with positive regulation of the cell cycle (de
229 Cárcer, Manning, and Malumbres 2011) and is regarded as an immediate-early response
230 gene (Winkles 1997) (Figure S17). Other examples include the genes JUN (C-Jun), FOS (c-
231 Fos), FOSB, and FOSL1 (FRA-1), that are also transiently over-expressed following release
232 from cell cycle arrest (Figure S13). The protein products of these genes are components
233 of the AP-1 transcription factor complex that is thought to be involved in cell
234 proliferation and cancer progression (Casalino et al. 2022; Milde-Langosch 2005).
235 Furthermore, the genes JUN, FOS, and FOSB are also members of the immediate-early

236 response gene family. The genes RELB, NFKB1, and NFKB2, are also transiently over-
237 expressed following release from cell cycle arrest (Figure S18). These genes are
238 members of the NF-kB family of transcription factors and are known to induce target
239 genes involved in initiation and progression of cancer (Costa et al. 2016; Tilborghs et al.
240 2017). We also observed that the apoptosis inhibitors BIRC2 and BIRC3 (Silke and Vaux
241 2001) are transiently over-expressed following release from cell cycle arrest (Figure
242 S27), whereas the gene BIRC5 (Survivin), which is a regulator of the mitotic cell cycle,
243 shows a periodic pattern of expression. Additional examples are shown in Figures S4-
244 S27.
245
246 The association of immediate-early response genes with topic k2 is not very surprising,
247 since these genes are known to be rapidly and transiently transcribed in response to
248 stimulation of cells by serum or growth factors such as EGF, subsequently promoting
249 phenotypic changes such as proliferation, differentiation, and survival (Winkles 1997). In
250 our study, it is likely that these genes facilitate cell-cycle reentry, promoting the
251 transition from the quiescent G0 phase into the G1 phase. The association of topic k2
252 with cervical cancer is less straightforward. One possible explanation for this association
253 is that, following release from cell cycle arrest, the cells enter a transient state lasting
254 approximately twelve hours (samples D1-D7), during which they activate molecular
255 mechanisms to resume the cell cycle. In normal tissue differentiation, cell cycle exit and
256 lineage-specification are often regulated by tissue-specific mechanisms (Theilgaard-
257 Mönch et al. 2022). It is plausible that cell cycle re-entry is similarly regulated by
258 mechanisms specific to HeLa cells and cervical cancer, the tissue from which they
259 originate.
260
261 For some of the 67 “core” cell cycle genes we observed that the spliced and unspliced
262 mRNA levels arrive at their first peak (either maximum or minimum) at the same time
263 (Figure 1B), and only after a few hours the typical time lag between the spliced and un-
264 spliced mRNA expression is acquired. This indicates that initially, during the few hours

265 after release from cell cycle arrest, the spliced and un-spliced transcripts are
266 synchronized in these genes. Additional evidence for this gradual acquisition of time lag
267 between spliced and un-spliced mRNA can be seen for larger sets of genes with periodic
268 behavior (Figure S2A-B). This indicates that in the transient cell state following release
269 from cell cycle arrest, the transcriptional timing mechanisms operate in a way that is
270 different from the normal cycling state. One possible explanation is that during this
271 period, the primary mRNA transcribed from these genes is spliced, processed, and
272 translated with minimal delay in order to rebuild the necessary mechanisms for cell
273 cycle re-entry as soon as possible. According to this hypothesis, splicing mechanisms
274 participate in the immediate-early response mechanism. Further studies at higher time
275 resolution are required to systematically test this hypothesis and inspect the
276 relationship between mRNA splicing and the immediate-early response gene program.
277

278 **ACKNOWLEDGEMENTS**

279 We wish to thank Shahar Alon and all the member of our lab for useful comments and
280 suggestions.
281

282 **DECLARATION OF INTEREST STATEMENT**

283 The authors have declared that no competing interests exist.
284

285 **AUTHOR CONTRIBUTIONS**

286 Study initiation and conception – T.M. and T.K.; Data analysis - T.M. and T.K.; Topic
287 modeling – T.M, J.G., and T.K.; Other intellectual contribution – Y.T.; Manuscript writing
288 – T.M. and T.K.
289

290 **FUNDING**

291 T.M., Y.T., and T.K. were supported by the Israel Science Foundation (ICORE no. 1902/12
292 and Grants no. 1634/13, 2017/13, and 1814/20), the Israel Ministry of Health (Grant no.
293 3-10146), the EU-FP7 (Marie Curie International Reintegration Grant no. 618592), the

Data Science Institute at Bar-Ilan University (seed grant), the ICRF (Grant no. 19-101-PG), the Israel Ministry of Science (Grant no. 3-16220), the Israel Ministry of Justice (Veadat Haezvonot), and the Israel Cancer Association (Grant no. 20240114). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

300 **REFERENCES**

301 Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation."
302 *Journal of Machine Learning Research* 3: 993–1022.

303 Carbonetto, Peter, Kevin Luo, Kushal Dey, Joyce Hsiao, and Matthew Stephens. 2021.
304 "fastTopics: Fast Algorithms for Fitting Topic Models and Non-Negative Matrix
305 Factorizations to Count Data. R Package Version 0.4-11."

306 de Cárcer, Guillermo, Gerard Manning, and Marcos Malumbres. 2011. "From Plk1 to
307 Plk5." *Cell Cycle* 10(14): 2255–62. doi:10.4161/cc.10.14.16494.

308 Casalino, Laura, Francesco Talotta, Amelia Cimmino, and Pasquale Verde. 2022. "The
309 Fra-1/AP-1 Oncoprotein: From the 'Undruggable' Transcription Factor to
310 Therapeutic Targeting." *Cancers* 14(6): 1480. doi:10.3390/cancers14061480.

311 Chen, Jing, Eric E. Bardes, Bruce J. Aronow, and Anil G. Jegga. 2009. "ToppGene Suite for
312 Gene List Enrichment Analysis and Candidate Gene Prioritization." *Nucleic Acids*
313 *Research*. doi:10.1093/nar/gkp427.

314 Chen, Jing, Huan Xu, Bruce J. Aronow, and Anil G. Jegga. 2007. "Improved Human
315 Disease Candidate Gene Prioritization Using Mouse Phenotype." *BMC*
316 *Bioinformatics* 8(1): 392. doi:10.1186/1471-2105-8-392.

317 Costa, Rui M. Gil Da, Margarida M. S. M. Bastos, Rui Medeiros, and Paula A. Oliveira.
318 2016. "The NFκB Signaling Pathway in Papillomavirus-Induced Lesions: Friend or
319 Foe?" *Anticancer Research* 36(5): 2073–83.

320 Dey, Kushal K., Chiaowen Joyce Hsiao, and Matthew Stephens. 2017. "Visualizing the
321 Structure of RNA-Seq Expression Data Using Grade of Membership Models."
322 *PLoS Genetics* 13(3): 1–23. doi:10.1371/journal.pgen.1006599.

323 Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali
324 Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR:
325 Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29(1): 15–21.
326 doi:10.1093/bioinformatics/bts635.

327 Dominguez, Daniel, Yi Hsuan Tsai, Nicholas Gomez, Deepak Kumar Jha, Ian Davis, and
328 Zefeng Wang. 2016. "A High-Resolution Transcriptome Map of Cell Cycle Reveals
329 Novel Connections between Periodic Genes and Cancer." *Cell Research* 26(8):
330 946–62. doi:10.1038/cr.2016.84.

331 Dominguez, Daniel, Yi Hsuan Tsai, Robert Weatheritt, Yang Wang, Benjamin J. Blencowe,
332 and Zefeng Wang. 2016. "An Extensive Program of Periodic Alternative Splicing
333 Linked to Cell Cycle Progression." *eLife* 5(MARCH2016): 1–19.
334 doi:10.7554/eLife.10288.

335 Herschman, H. R. 1991. "Primary Response Genes Induced by Growth Factors and
336 Tumor Promoters." *Annual Review of Biochemistry* 60: 281–319.
337 doi:10.1146/annurev.bi.60.070191.001433.

338 Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of
339 Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome biology*
340 15(12): 550. doi:10.1186/s13059-014-0550-8.

341 La Manno, Gioele, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner,
342 Viktor Petukhov, Katja Lidschreiber, et al. 2018. "RNA Velocity of Single Cells."
343 *Nature* 560(7719): 494–98. doi:10.1038/s41586-018-0414-6.

344 Milde-Langosch, Karin. 2005. "The Fos Family of Transcription Factors and Their Role in
345 Tumourigenesis." *European Journal of Cancer* 41(16): 2449–61.
346 doi:10.1016/j.ejca.2005.08.008.

347 O'Donnell, Amanda, Zaneta Odrowaz, and Andrew D. Sharrocks. 2012. "Immediate-Early
348 Gene Activation by the MAPK Pathways: What Do and Don't We Know?"
349 *Biochemical Society Transactions* 40(1): 58–66. doi:10.1042/BST20110636.

350 Silke, John, and David L. Vaux. 2001. "Two Kinds of BIR-Containing Protein - Inhibitors of
351 Apoptosis, or Required for Mitosis." *Journal of Cell Science* 114(10): 1821–27.
352 doi:10.1242/jcs.114.10.1821.

353 Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L.
354 Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment
355 Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide
356 Expression Profiles." *Proceedings of the National Academy of Sciences* 102(43):
357 15545–50. doi:10.1073/pnas.0506580102.

358 Theilgaard-Mönch, Kim, Sachin Pundhir, Kristian Reckzeh, Jinyu Su, Marta Tapia,
359 Benjamin Furtwängler, Johan Jendholm, et al. 2022. "Transcription Factor-Driven
360 Coordination of Cell Cycle Exit and Lineage-Specification in Vivo during
361 Granulocytic Differentiation." *Nature Communications* 13(1): 3595.
362 doi:10.1038/s41467-022-31332-1.

363 Tilborghs, Sam, Jerome Corthouts, Yannick Verhoeven, David Arias, Christian Rolfo, Xuan
364 Bich Trinh, and Peter A. van Dam. 2017. "The Role of Nuclear Factor-Kappa B
365 Signaling in Human Cervical Cancer." *Critical Reviews in Oncology/Hematology*
366 120: 141–50. doi:10.1016/j.critrevonc.2017.11.001.

367 Tullai, John W., Michael E. Schaffer, Steven Mullenbrock, Gabriel Sholder, Simon Kasif,
368 and Geoffrey M. Cooper. 2007. "Immediate-Early and Delayed Primary Response
369 Genes Are Distinct in Function and Genomic Architecture." *Journal of Biological*
370 *Chemistry* 282(33): 23981–95. doi:10.1074/jbc.M702044200.

371 Winkles, Jeffrey A. 1997. "Serum- and Polypeptide Growth Factor-Inducible Gene
372 Expression in Mouse Fibroblasts." In *Progress in Nucleic Acid Research and*
373 *Molecular Biology*, ed. Kivie Moldave. Academic Press, 41–78.
374 doi:10.1016/S0079-6603(08)60033-1.

375

376 **APPENDICES**

377 **Supplementary information:** Supplementary text and figures.

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

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

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

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

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

384 **Table S6:** A list of immediate-early response genes from Tullai *et al*.

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

387

FIGURES:

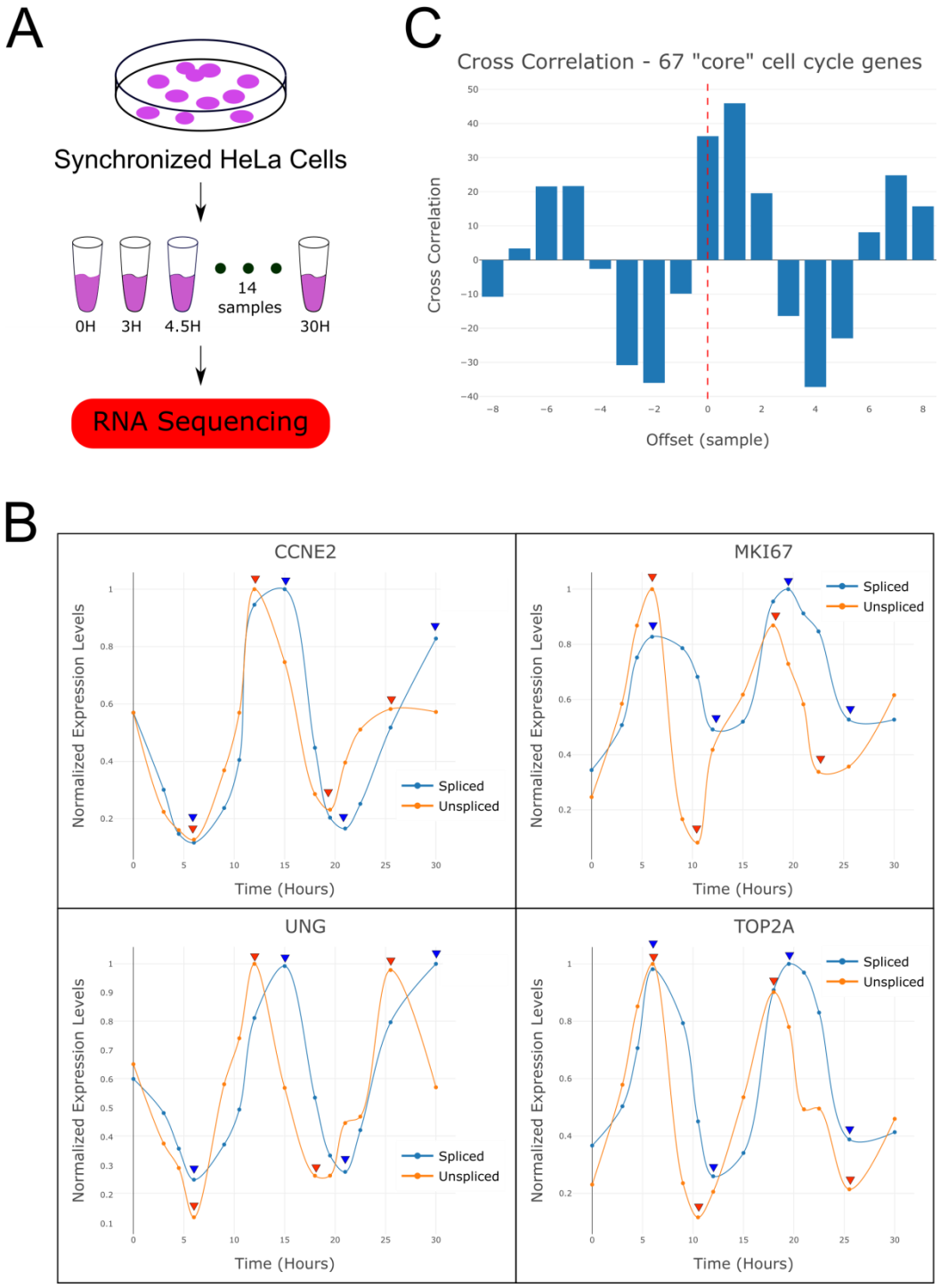


Figure 1: RNA velocity analysis of periodically expressed genes reveals a time lag between spliced and un-spliced mRNA.

393 (A) Shown is a sketch of the experiment performed by Dominguez *et al.* from which the
394 RNAseq dataset was produced. HeLa cells were sampled at 14 consecutive time points
395 following release from cell cycle arrest and their RNA was sequenced. (B) Shown are
396 expression levels over time of spliced and unspliced mRNA for selected genes that are
397 known to be over-expressed at the G1-S phases (CCNE2, UNG) or the G2-M phases
398 (MKI67, TOP2A) of the cell cycle. These indicate that the cells completed approximately
399 two cell divisions during the experiment. It can be seen that the yet-unspliced mRNA
400 again precedes the spliced mRNA, as expected. Expression levels were normalized to
401 their maximum and spline curves were plotted through the data points to assist
402 visibility. (C) A cross-correlation plot between the spliced and un-spliced mRNA levels
403 averaged over the 67 "core" cell cycle genes (i.e. periodically expressed genes identified
404 by Dominguez *et al.*) shows a significant time lag between the spliced and yet un-spliced
405 mRNA.
406

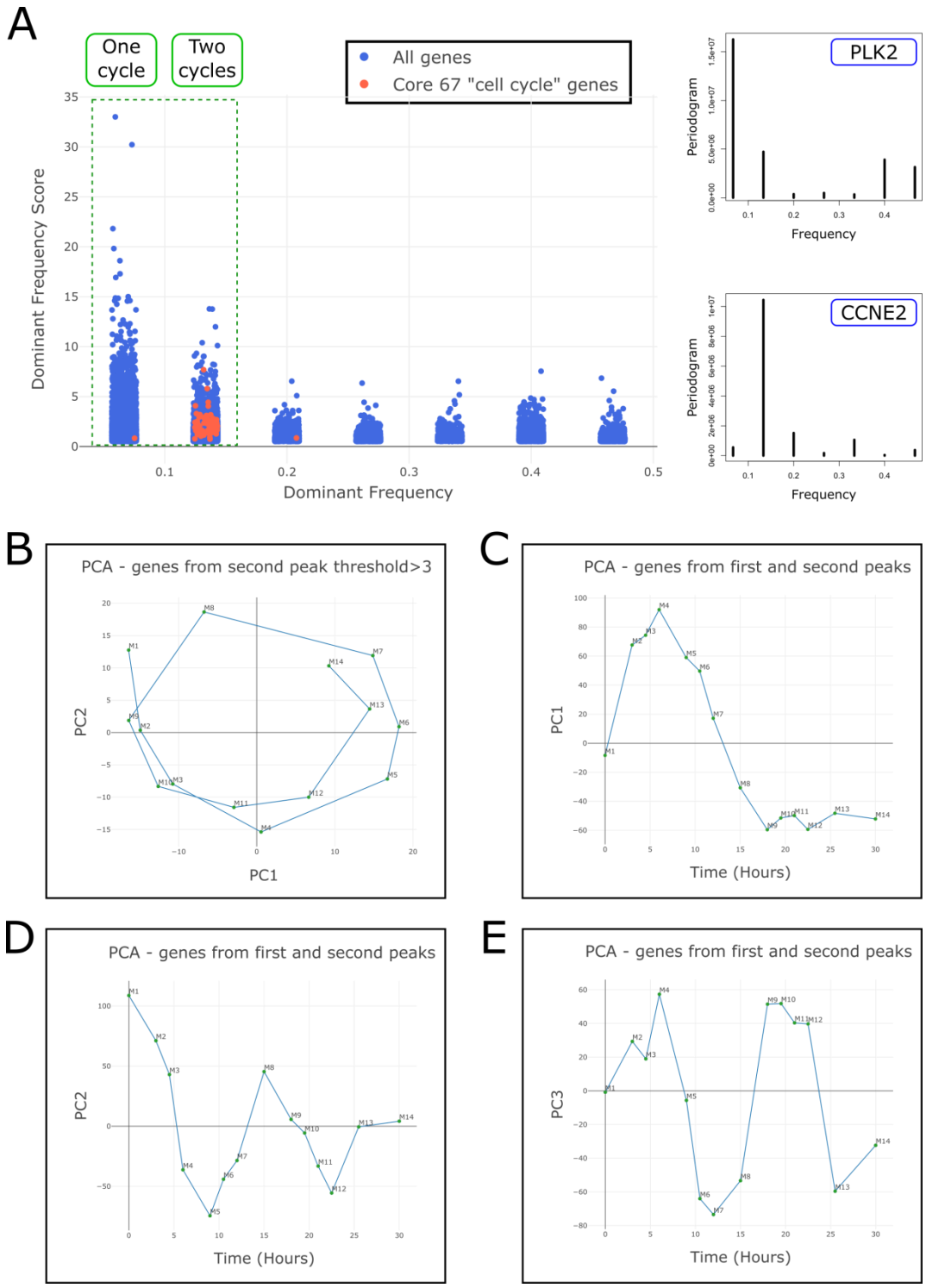


Figure 2: Fourier analysis identifies sets of genes with potentially transient and oscillatory behaviors over time.

(A) For each gene, a periodogram was calculated (see specific examples for PLK2 and CCNE2 in the right panels) and its "dominant frequency" and "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 scores that are higher than those within other dominant frequencies, whereas genes in the 3rd dominant frequency and onwards have scores that are similar to those of randomly shuffled datasets (Figure S1). This indicates 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 completes almost two 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 a transient behavior along PC1 axis and periodic behaviors along the PC2 and PC3 axes (that also create a circular trajectory in the PC2 vs. PC3 plane).

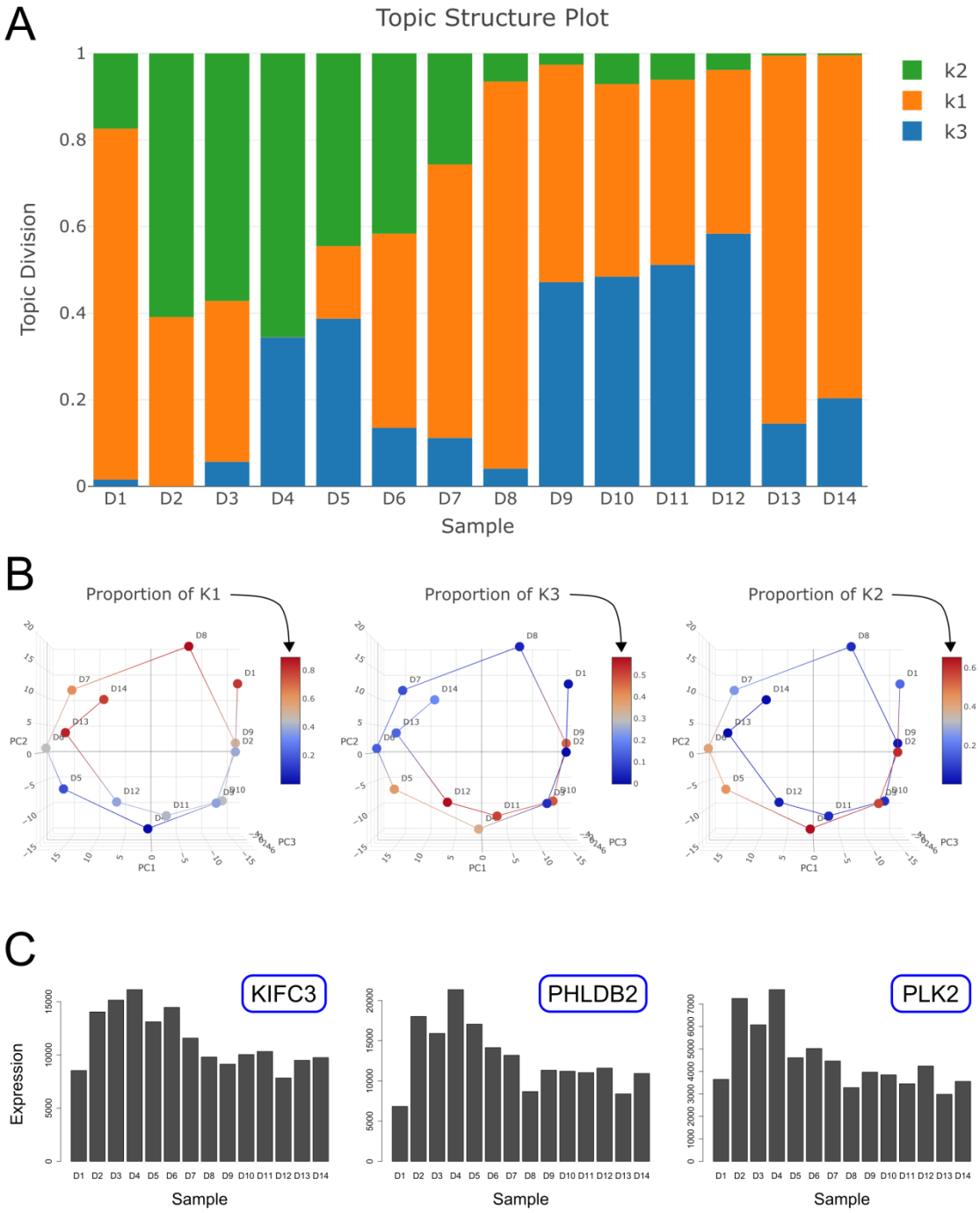
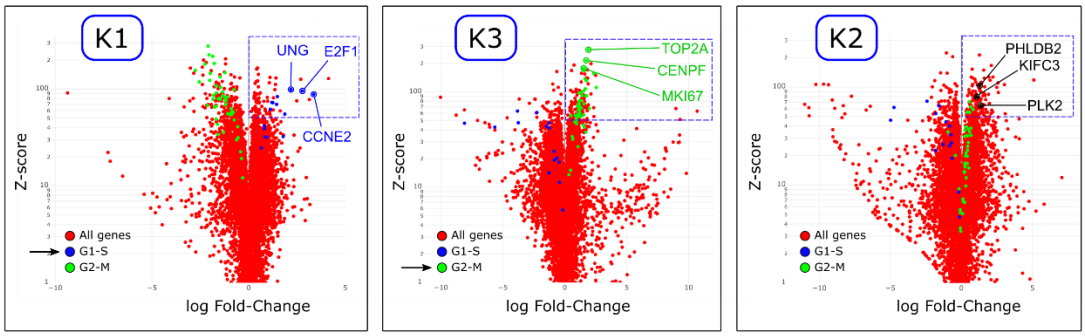


Figure 3: Topic modeling reveals three temporal components with periodic and transient behaviors over time.

(A) Shown is a structure plot of the proportions of each component/topic within each sample. Topics k1 and k3 are periodically expressed over the course of both cycles (samples D1-D14), whereas topic k2 is transiently over-expressed mainly during the first

433 cycle (samples D1-D7) following release from cell cycle arrest. (B) Shown is a
434 visualization of the proportions of each topic along the circular trajectory in latent space
435 representing the two cell cycles (red – high proportion, blue – low proportion). It can be
436 seen that topics k1 and k3 are over-expressed in distinct phases during both cycles,
437 whereas topic k2 is over-expressed during the first cycle only. Note that topic k1 is
438 expressed mainly in the upper half of the circular trajectory, whereas topic k3 is
439 expressed mainly in its lower half. (C) Shown are the expression levels of three
440 representative genes that are associated with topic k2 and are transiently over-
441 expressed following release from cell cycle arrest.
442

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

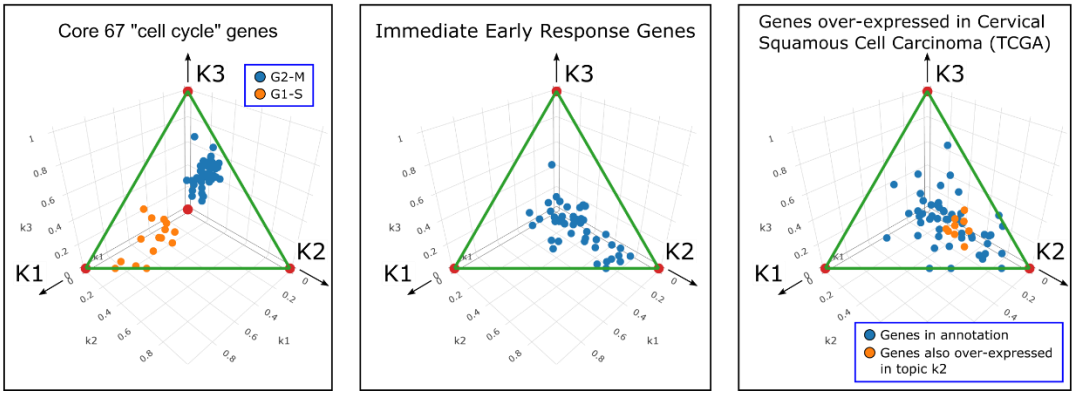
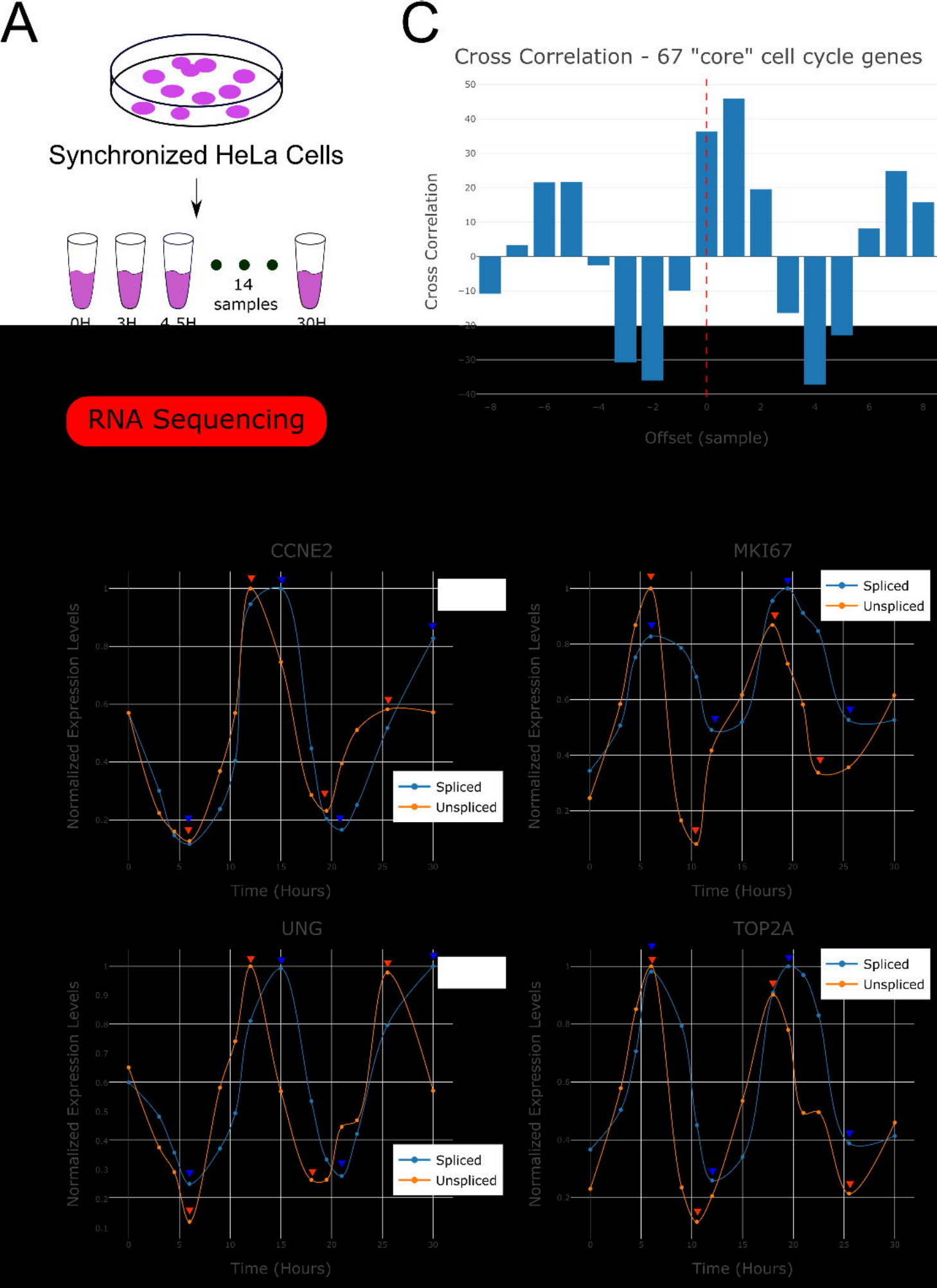
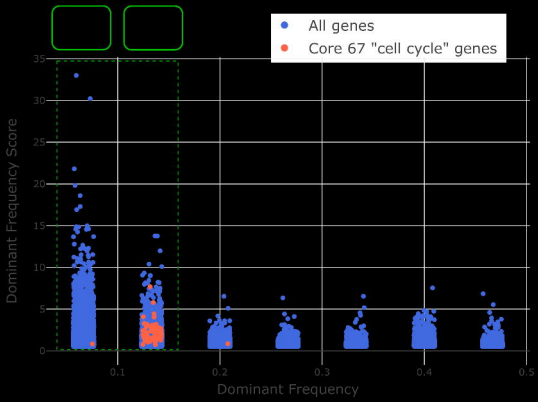


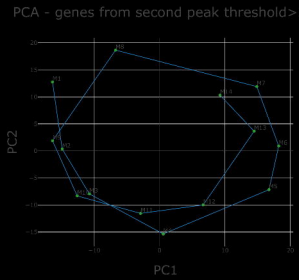
Figure 4: Gene Ontology (GO) enrichment shows that the two periodic components correspond to the G1-S and G2-M phases of the cell cycle, whereas the third transient component is related to immediate-early response, regulation of cell proliferation, and cervical cancer.

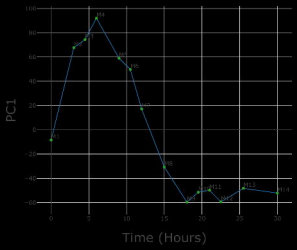
(A) Shown are volcano plots illustrating the over-expression of each gene in each of the three topics. It can be seen that genes known to be over-expressed in the G1-S phases of the cell cycle are over-expressed in topic k1 (left panel), genes known to be over-expressed in the G2-M phases of the cell cycle are over-expressed in topic k3 (middle panel), and genes that are transiently over-expressed following release from cell cycle

453 arrest are over-expressed in topic k2 (right panel). (B) GO enrichment analysis of genes
454 that are over-expressed in topic k2 (Z-score>50) shows many of them are related to
455 cervical cancer or involved in response to growth factors (Table S4). (C) Shown are the
456 posterior probabilities $p(topic|gene)$ for selected gene sets or specific GO terms. Each
457 point represents a gene, and its components along the k1, k2, and k3 axes are the
458 posterior probabilities $p(k1|gene)$, $p(k2|gene)$, and $p(k3|gene)$, which represent the
459 association between that gene and each of the three topics. It can be seen that topic k1
460 is highly associated with genes that are known to be over-expressed in the G1-S phases
461 of the cell cycle (left panel) and that topic k3 is highly associated with genes that are
462 known to be over-expressed in the G2-M phases of the cell cycle. Topic k2 is associated
463 with genes that are related to the immediate-early response program and to cervical
464 carcinoma (middle and right panels). Note that all points (=genes) are located on the
465 triangle-shaped two-dimensional simplex since the sum of the posterior probabilities for
466 each gene equals 1.
467


A
Synchronized HeLa Cells
14
samples
0H
3H
4.5H
30H
RNA Sequencing
C
Cross Correlation - 67 "core" cell cycle genes
Cross Correlation
Offset (sample)
CCNE2
MKI67
UNG
TOP2A
Normalized Expression Levels
Time (Hours)
Spliced
Unspliced

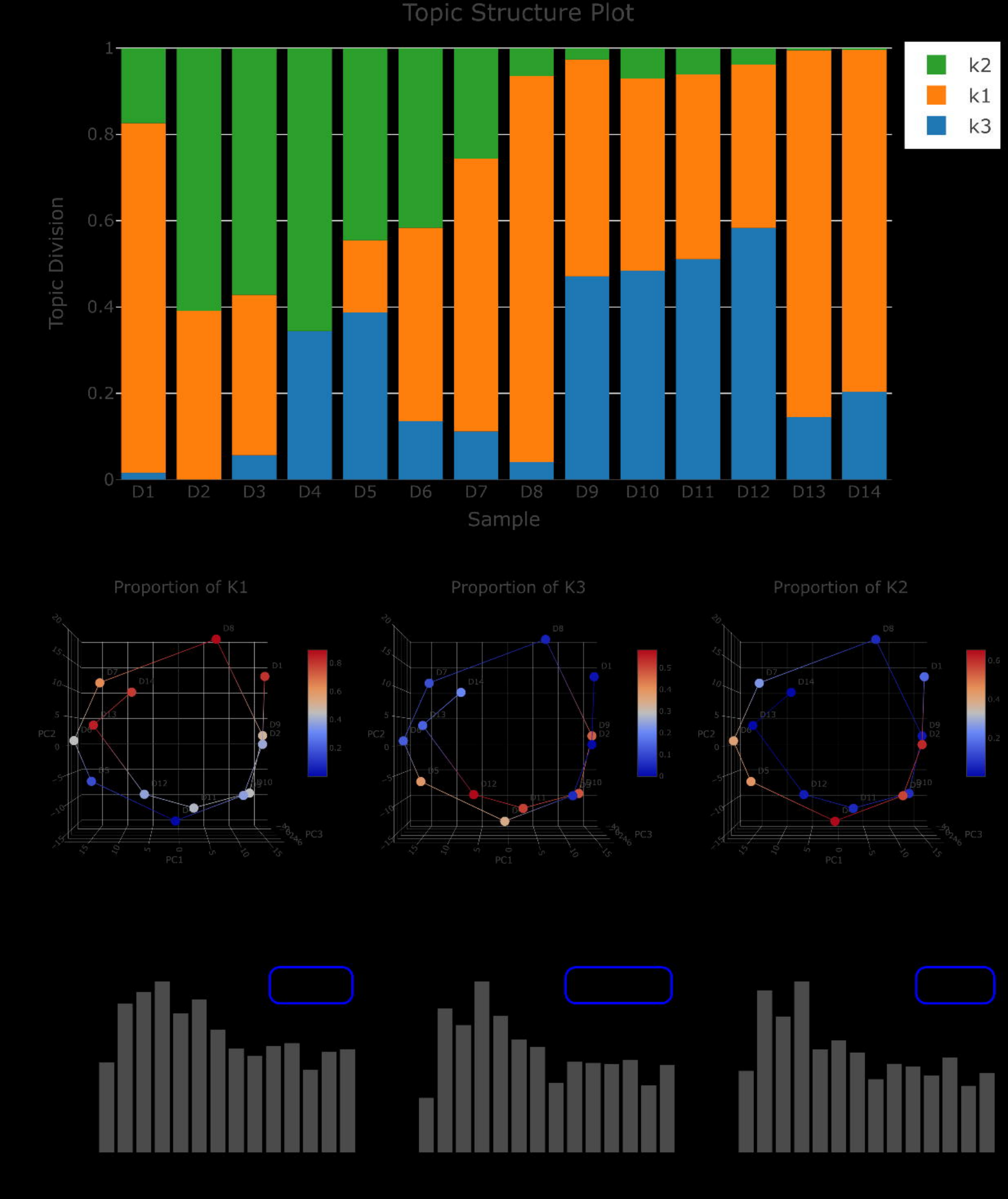

All genes
Core 67 "cell cycle" genes
Dominant Frequency Score
Dominant Frequency
0
5
10
15
20
25
30
35
0.1
0.2
0.3
0.4
0.5

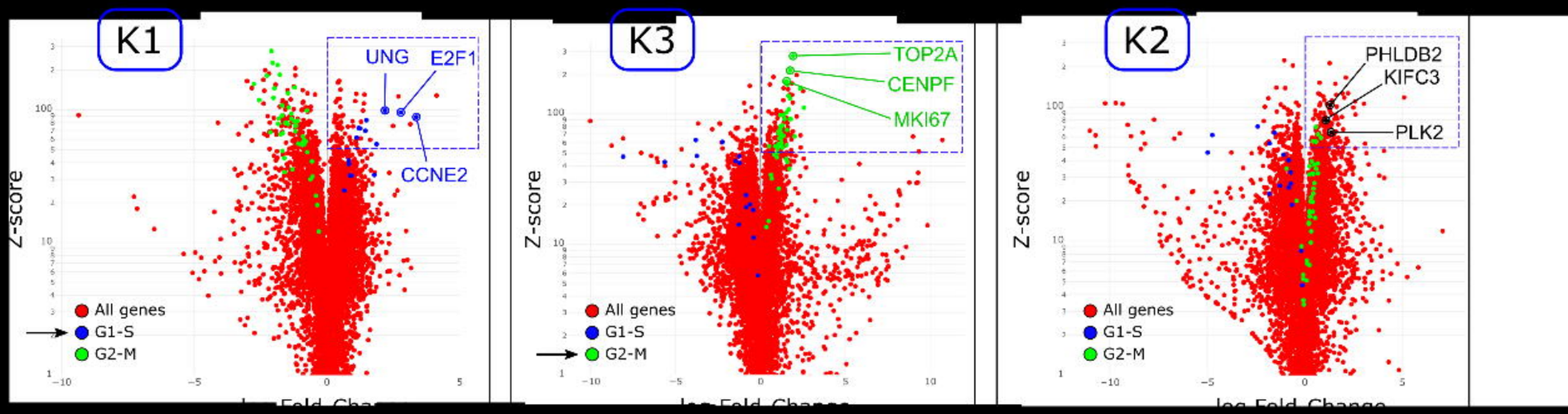
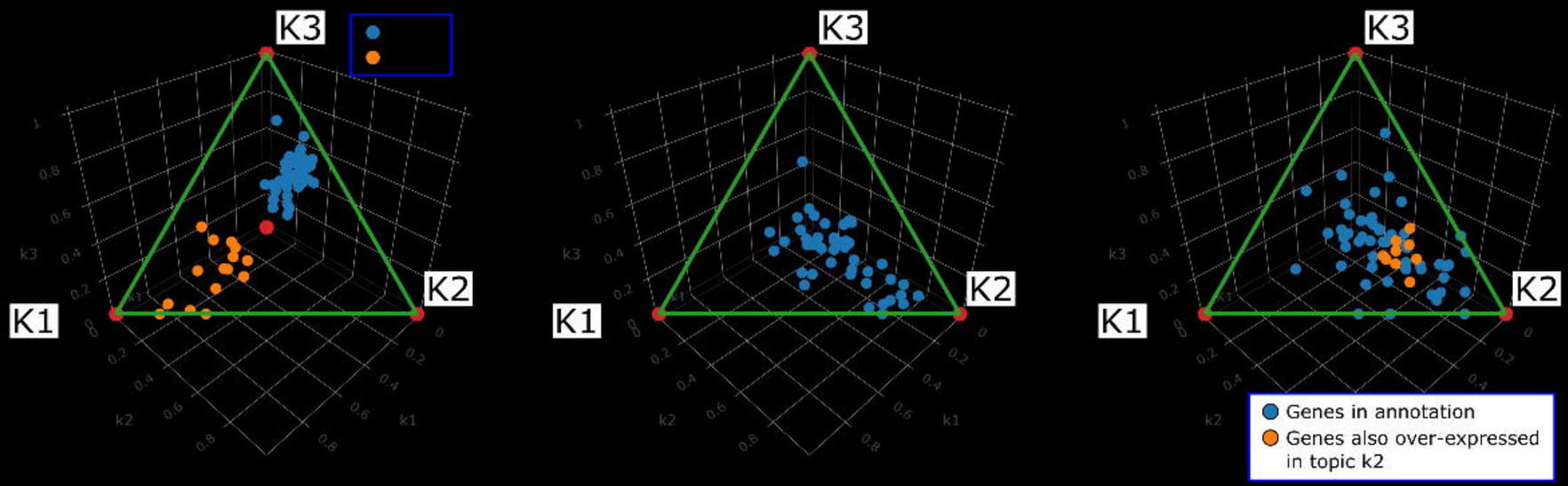

PCA - genes from second peak threshold>3
PC2
PC1


PCA - genes from first and second peaks
PC1
Time (Hours)

PCA - genes from first and second peaks
PC2
Time (Hours)

PCA - genes from first and second peaks
PC3
Time (Hours)


Topic Structure Plot
Topic Division
Sample
k2
k1
k3
D1
D2
D3
D4
D5
D6
D7
D8
D9
D10
D11
D12
D13
D14
Proportion of K1
Proportion of K3
Proportion of K2
PC1
PC2
PC3



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