Unbiased Boolean analysis of public gene expression data for cell cycle gene identification

Sarah A. Dabydeen[a], Arshad Desai[b,c], and Debashis Sahoo[a,d,e,*]
[a]Department of Pediatrics, [c]Department of Cellular and Molecular Medicine, [d]Department of Computer Science and Engineering, and [e]Moores Cancer Center, University of California, San Diego, La Jolla, CA 92093; [b]Ludwig Institute for Cancer Research, La Jolla, CA 92093

ABSTRACT **Cell proliferation is essential for the development and maintenance of all organisms and is dysregulated in cancer. Using synchronized cells progressing through the cell cycle, pioneering microarray studies defined cell cycle genes based on cyclic variation in their expression. However, the concordance of the small number of synchronized cell studies has been limited, leading to discrepancies in definition of the transcriptionally regulated set of cell cycle genes within and between species. Here we present an informatics approach based on Boolean logic to identify cell cycle genes. This approach used the vast array of publicly available gene expression data sets to query similarity to *CCNB1*, which encodes the cyclin subunit of the Cdk1-cyclin B complex that triggers the G2-to-M transition. In addition to highlighting conservation of cell cycle genes across large evolutionary distances, this approach identified contexts where well-studied genes known to act during the cell cycle are expressed and potentially acting in nondivision contexts. An accessible web platform enables a detailed exploration of the cell cycle gene lists generated using the Boolean logic approach. The methods employed are straightforward to extend to processes other than the cell cycle.**

Monitoring Editor
Francis A. Barr
University of Oxford

Received: Jan 7, 2019
Revised: Apr 4, 2019
Accepted: May 29, 2019

INTRODUCTION

The cell cycle has been extensively investigated using diverse approaches in different experimental models. Seminal breakthroughs in the cell cycle field came from genetic and biochemical analyses in models such as yeasts, invertebrate and vertebrate eggs/early embryos, and human tissue culture cells. These efforts have defined a large set of components that execute the many complex events in the cell cycle (Morgan, 2006).

The development of microarray technology in the late 1990s spurred efforts to employ transcriptional coregulation as an unbiased means to define genes whose expression varies in coordination with progression through the cell cycle (Cho *et al.*, 1998; Whitfield *et al.*, 2002). Additional targeted studies have revealed involvement of RB-E2F, DREAM, and MMB-FOXM1 transcriptional regulatory complexes in cell cycle regulation in mammals (Wen *et al.*, 2008; Lewis *et al.*, 2012; Sim *et al.*, 2012; Bertoli *et al.*, 2013; DeBruhl *et al.*, 2013; Sadasivam and DeCaprio, 2013; Fischer *et al.*, 2016). Previous attempts of the meta-analysis of different synchronization-based transcriptional data sets have led to conflicting results (de Lichtenberg *et al.*, 2005; Marguerat *et al.*, 2006; Gauthier *et al.*, 2008; Wang *et al.*, 2016; Giotti *et al.*, 2017). A currently available website resource lists 378 human cell cycle–associated genes (Gauthier *et al.*, 2008, 2010; Santos *et al.*, 2015). A recent meta-analysis of these data sets has also proposed a list of 1419 cell cycle genes (Giotti *et al.*, 2017). Notably, there are only 165 common genes between the above two different lists of candidate cell cycle genes.

The use of an experimental perturbation (synchronization-release coupled with time-sampling of expression) to identify genes with periodic variation in expression during the cell cycle has generated five human gene expression data sets (total 305 samples) that have been the focus of the majority of transcriptional regulation-guided cell cycle gene analysis to date. In contrast to this small number of synchronized cell data sets, microarray expression profiles of 25,955 human samples are publicly accessible. We have therefore focused on developing bioinformatic approaches to efficiently mine this large volume of public data to define cell cycle genes based on their transcriptional regulation. In our approach, there is no assumption that cell cycle genes exhibit periodic gene expression. Instead, the

This article was published online ahead of print in MBoC in Press (http://www.molbiolcell.org/cgi/doi/10.1091/mbc.E19-01-0013) on May 15, 2019.
*Address correspondence to: Debashis Sahoo (dsahoo@ucsd.edu).
Abbreviations used: BECC, Boolean Equivalent Correlated Clusters; GEO, Gene Expression Omnibus; NCBI, National Center for Biotechnology Information.

approach is based on Boolean logic applied to the very large number of available expression data sets. In addition to defining cell cycle genes in individual species and their conservation across species, we have built a website resource that facilitates analysis of transcriptional relationships for cell cycle genes in four commonly studied multicellular model organisms.

RESULTS

A Boolean equivalence approach based on public gene expression data sets identifies cell cycle genes in humans

The informatic approach that we employed to mine publicly available gene expression data sets (n = 25,955 human samples) is known as Boolean Equivalent Correlated Clusters (BECC; Figure 1A, Supplemental Figure S1). BECC compares the normalized expression of two genes across all data sets by searching for two sparsely populated, diagonally opposite quadrants out of four possible quadrants (high–low and low–high), employing the BooleanNet algorithm (Sahoo *et al.*, 2008). There are six potential gene relationships assessed by BooleanNet: two symmetric (Equivalent and Opposite) and four asymmetric (Sahoo *et al.*, 2008). Two genes are considered Boolean equivalent if they are positively correlated with only high–high and low–low gene expression values. Two genes are considered Boolean Opposite if they are negatively correlated with only high–low and low–high gene expression values. Asymmetric Boolean implications result when there is only one sparsely populated quadrant. The BECC algorithm only focuses on Boolean equivalent relationships to identify potentially functionally related gene sets.

To identify potential cell cycle genes with this approach, we employed BECC using *CCNB1* (which encodes cyclin B1) as a seed gene. Cyclin B1 is the binding partner of the kinase Cdk1 and

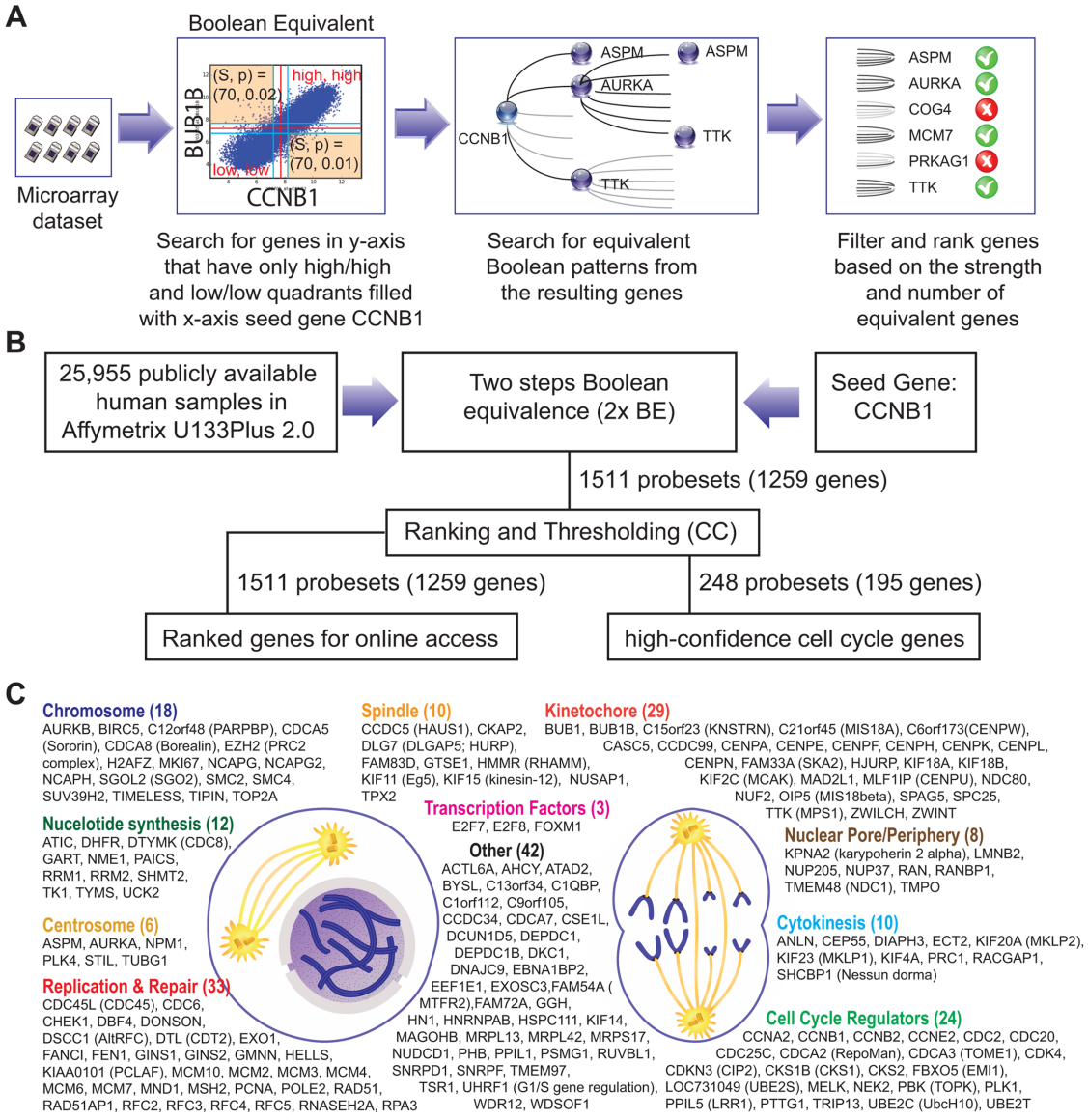

FIGURE 1: Approach and analysis in human. (A) Schematic algorithm of BECC that was performed on large gene expression data sets. (B) Individual steps of the BECC algorithm. First step is to perform Boolean analysis by following the Boolean equivalent relationship twice. The second step is to compute a score for each gene and rank the genes. Finally, a threshold is imposed using the StepMiner algorithm to define a list of high-confidence genes. (C) List of 195 high-confidence cell cycle genes in human categorized by their known biological function.

Cdk1-Cylin B1 complex activity defines the mitotic phase of the cell cycle throughout eukaryotes. All gene pairs with *CCNB1* were analyzed in a pooled human microarray data set built from 25,955 human samples, and each tested probeset on the array was ranked based on the strength of its correlation for an equivalent Boolean relationship with the *CCNB1* probeset. The data set exclusively uses the Affymetrix microarray platform where a group of short probes (probeset) target a particular gene at the 3′ end of the corresponding mRNA sequences. Each array contains multiple probesets for a gene for which the signal intensity and noise characteristics can vary significantly. In our analysis, we used probeset ID 214710_s_at as the seed *CCNB1* probe based on its robust signal and ability to capture a large dynamic range of expression. We note that a second probeset (228729_at) behaves very similarly and would yield essentially identical results.

The BECC algorithm was first used to identify a set of 145 probesets (ProbeSet A) that were Boolean equivalent to the *CCNB1* probe (Supplemental Figures S1 and S2). Then, the same algorithm was used to identify additional probesets that were Boolean equivalent to ProbeSet A; pooling the hits in the second step together with those in ProbeSet A resulted in ProbeSet B composed of 1511 probesets. BECC computes Boolean equivalences for two steps because any additional steps have the potential to add significant noise. All probesets in ProbeSet B were then comprehensively analyzed relative to each other to assess the strength of their equivalences. A Boolean equivalence score for each probeset within ProbeSet B was computed based on the weighted average of the correlation coefficient and slope in pairwise analysis with all other probesets, as described in *Materials and Methods*. This effort resulted in a ranked list of 1511 probesets, corresponding to 1259 unique genes, based on similarity to CCNB1. The entire ranked list of genes can be accessed online using our web resource. StepMiner, an algorithm that fits a step function to identify abrupt transitions in series data, was used to compute a threshold on the BE score to identify high-confidence cell cycle genes. Imposition of the threshold resulted in the identification of 248 significant probesets, representing 195 unique genes (Figures 1B and 2A, Supplemental Figures S1 and S2). These 195 genes represent high-confidence hits from the ranked list generated prior to threshold imposition.

The 195 high-confidence genes identified by this unbiased data mining approach could be categorized into the following groups based on the literature (Figure 1C): cytokinesis (n = 10), centrosome (n = 6), kinetochore (n = 29), spindle (n = 10), chromosome (n = 18), cell cycle regulators (n = 24), replication and repair (n = 33), nuclear pore/periphery (n = 8), nucleotide synthesis (n = 12), transcription factors (n = 3), and other (n = 42). Gene ontology (Ashburner *et al.*, 2000), MSigDB pathway (Subramanian *et al.*, 2005), Reactome (Fabregat *et al.*, 2018), and DAVID gene functional classification tools (Dennis *et al.*, 2003; Hosack *et al.*, 2003) revealed enrichment of the term Cell Cycle in the list of 195 genes (Supplemental Figure S7). It is important to note that in a Boolean equivalence analysis the final list of predicted genes is not very sensitive to the choice of the initial seed gene within a related set—if a gene other than *CCNB1* from the list of 195 genes was employed as a seed gene, the top ranking genes will always be present. For example, 155 high-confidence cell cycle genes were predicted using *CCNB2* as a seed gene instead of *CCNB1*, and 154 genes of these 155 genes overlap with the 195 genes identified using *CCNB1* as a seed. *CCNB2* is also expressed in dividing cells but unlike *CCNB1* is dispensable for viability (Brandeis *et al.*, 1998).

Comparison of the cell cycle gene set defined by Boolean equivalence analysis to genes identified based on periodic expression in a single cell cycle

We next compared the 195 high-confidence gene list to six previous studies that identified periodic gene expression patterns in synchronized cell cycle experiments (Cho *et al.*, 1998; Whitfield *et al.*, 2002; Bar-Joseph *et al.*, 2008; Grant *et al.*, 2013; Peña-Diaz *et al.*, 2013; Giotti *et al.*, 2017). In the Cho *et al.* (1998) study, expression of only 80 of the 195 genes was measured and 32 (40%) of these 80 genes were defined as being cell cycle–regulated. In the Whitfield *et al.* (2002) study, 172 of the 195 genes were measured, and 106 (62%) were defined as cell cycle–regulated. Similarly, the overlap with other studies are as follows: Bar-Joseph *et al.* (2008) (108/167, 65%), Grant *et al.* (2013) (99/172, 58%), Peña-Diaz *et al.* (2013) (97/195, 50%), and Giotti *et al.* (2017) (127/167, 76%). Figure 2A presents the heatmaps for periodic expression for the genes we identified that overlap with the ones analyzed by Whitfield *et al.* (2002) (172 genes, 314 clones) and Bar-Joseph *et al.* (2008) (167 genes, 262 probesets).

We note that the concordance between the different periodic gene expression studies is limited, for example, from our 195 gene set, only 21 were found in all six studies and 38 were not found in any of the studies. While these differences may be related to technical reasons, the Boolean approach, which does not rely on artificial synchronization-release in culture, offers a complementary approach to defining cell cycle–regulated genes. Among the list of 195 genes, two candidate cell cycle genes, *ATAD2* (Figure 2B) (Whitfield *et al.*, 2002; Bar-Joseph *et al.*, 2008; Grant *et al.*, 2013; Giotti *et al.*, 2017) and *CDCA7* (Figure 2C) (Whitfield *et al.*, 2002; Peña-Diaz *et al.*, 2013), are highly ranked but have been subject to limited characterization. We also note that certain standard cell cycle genes employed in the synchronization-release analysis in culture, such as the gene encoding *VEGF-C* (Whitfield *et al.*, 2002), are not Boolean equivalent to *CCNB1* (Figure 2D).

The cell division cycle is a fundamental biological process involving genes that are likely to be essential for viability. CRISPR/Cas9 technology has been recently adapted to perform large-scale gene essentiality screens in human cultured cells (Blomen *et al.*, 2015; Hart *et al.*, 2015; Wang *et al.*, 2015; Bertomeu *et al.*, 2018). We plotted the 195 high-confidence cell cycle genes that were designated as being essential in 10 different cell lines (Figure 2E). We found that 46 genes (24%) were essential in all 10 cell lines, 62 genes (32%) were categorized as being not essential, and the remainder showed context-dependent essentiality, in that they were essential in more than one cell line but not in all cell lines. Thus 133 (68%) of the genes identified by BECC have been designated as essential in more than one genome-wide screen, indicating a strong enrichment for essential genes in the gene set defined by the Boolean equivalence approach.

In relation to the complementary nature of the Boolean equivalence approach relative to the synchronization-release experiments in culture, the expression of two cell cycle genes may be out-of-phase with each other in a single cell cycle but still identified as equivalent by BECC. An example of an out-of-phase expression is shown for *CCNB1* and *CCNE2* (Figure 2, F and G, Supplemental Figure S4). Despite the striking negative correlation across a single cell cycle, *CCNE2* is identified by BECC as a strong Boolean equivalent gene with *CCNB1* (Figure 2H). This is likely because the majority of expression data sets are from tissue samples with varying numbers of dividing cells at different phases, resulting in average expression profiles across all cell cycle phases. Thus, the expression value of genes that is specifically expressed in dividing cells will be proportional to the fraction of dividing cells in the tissue/sample, resulting in Boolean equivalence, despite being out-of-phase in a

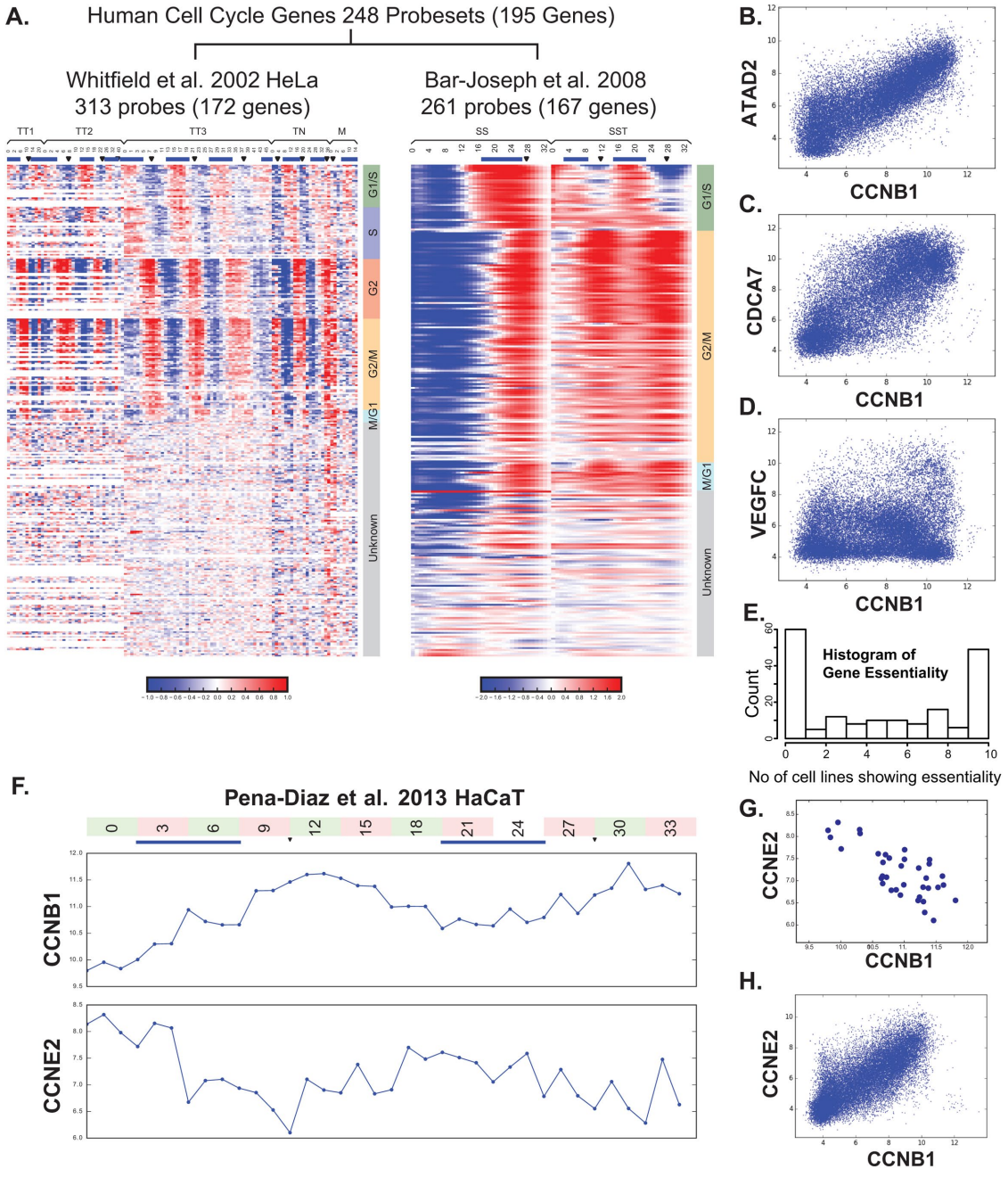

FIGURE 2: Comparison with human synchronized cell analysis. (A) The heatmap shows that the majority of BECC-identified high-confidence cell cycle genes are periodically expressed at different phases of the cell cycle in two synchronized cell cycle experiments. (B–D, H) Scatterplots of *ATAD2*, *CDCA7*, *VEGFC*, and *CCNE2* with *CCNB1*, respectively, in the public human data set with 25,955 samples. (E) Histogram of cell lines where the genes show essentiality. (F, G) Expression patterns of *CCNB1* (G2/M gene) and *CCNE2* (G1/S gene) in Peña-Diaz *et al.* (2013). HaCaT data set shows strong negative correlation. (H) *CCNB1* and *CCNE2* show strong positive correlation.

single cell cycle. This line of reasoning also explains why seeding the BECC algorithm with *CCNB1*, whose expression is regulated to occur after S-phase, identified genes acting in early S-phase. Thus, the Boolean equivalence approach, by mining public data sets with a single seed gene, is able to identify cell cycle genes, even if they are expressed at different phases in a single cell cycle.

To assess whether any of the 38 high-confidence cell cycle genes identified by our approach, but not in the prior synchronized cell culture studies, exhibit cell cycle–regulated expression, we turned to a new single-cell RNASeq data set GSE121265 (Hsiao *et al.*, 2019). This data set used iPSC lines that were genetically engineered to express the FUCCI (fluorescent ubiquitination cell cycle indicator) reporters to indicate cell cycle status. Pearson's correlation coefficient was computed between *CCNB1*, the green Fucci reporter (S/G2/M), and the 38 genes (Supplemental Figure S5A). The S/G2/M reporter and *CDC20* were strongly correlated with *CCNB1* (correlation coefficient 0.71 with EGFP Fucci signal and 0.79 with CDC20; Supplemental Figure S5, B and C). CCNB1 was poorly correlated with NANOG (correlation coefficient 0.31; Supplemental Figure S5D), which is not considered a cell cycle gene. Notably,

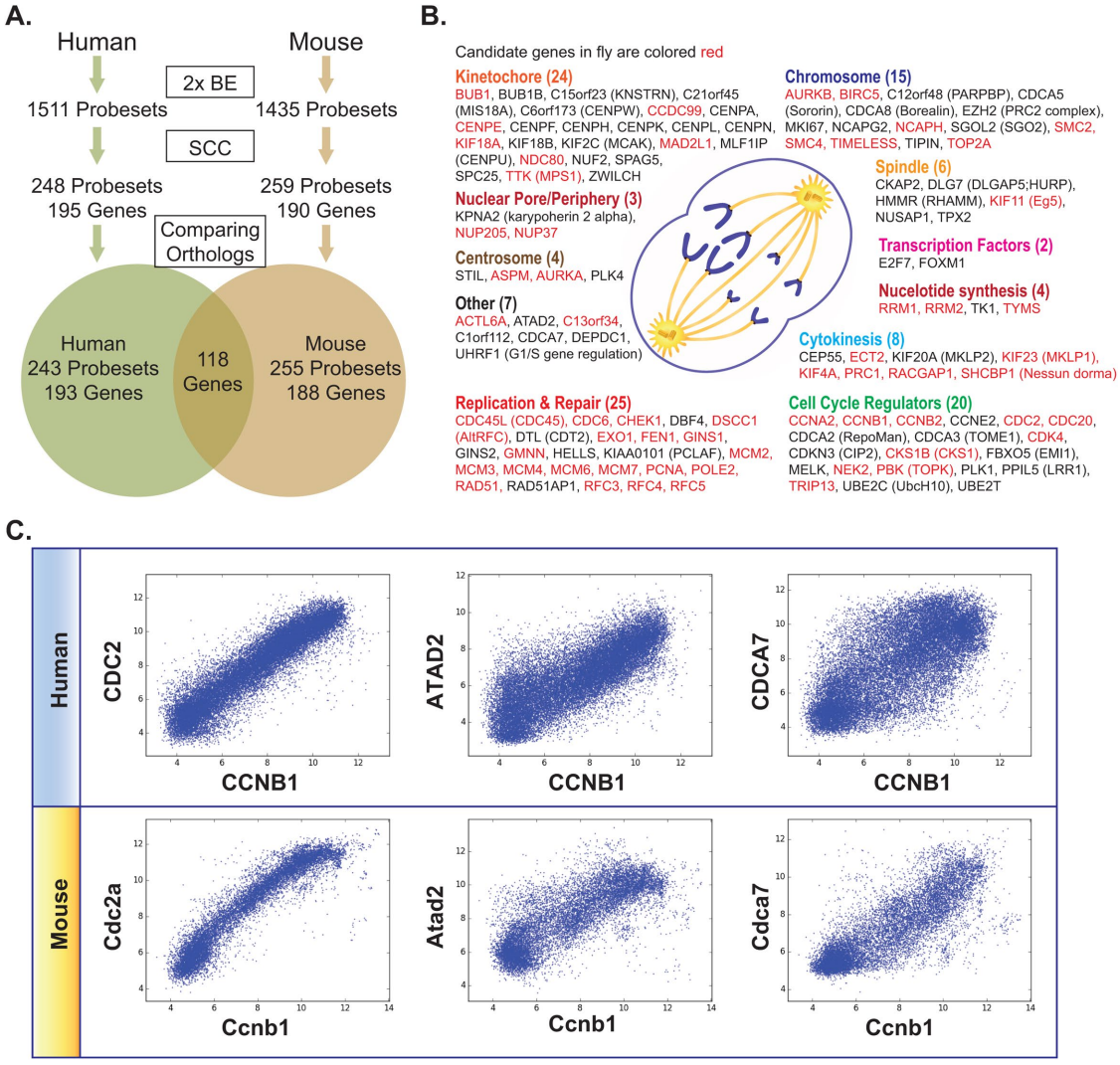

FIGURE 3: Comparison of high-confidence gene lists between human and mouse. (A) Identification of candidate cell cycle genes in human and mouse using BECC algorithm; comparison of orthologous genes between humans and mice; Venn diagram. (B) List of 103 genes conserved between humans and mice. Red color highlights genes whose expression patterns are also conserved in fly data sets. (C) Three candidate cell cycle genes (*CDC2*, mouse orthologue *Cdc2a*, *ATAD2*, and *CDCA7*) show strong positive correlation with CCNB1 in publicly available pooled human ($n = 25,955$) and mouse ($n = 11,758$) data sets. *CDC2* is a highly studied cell cycle gene, whereas *ATAD2* and *CDCA7* have been subjected to only limited analysis.

28 genes from the list of 38 genes exhibited correlation coefficients greater than 0.6 with CCNB1 (Supplemental Figure S5; for specific examples, see scatterplots in Supplemental Figure S5, E and F).

Assessing conservation of the cell cycle gene set between humans and mice

Mice are the premier mammalian genetic model and share a common ancestor with humans ~80 million years ago. Genetic mouse models are critical for understanding normal and pathological cell cycle regulation, leading us to conduct the same analysis reported above in humans with 11,758 publicly available mouse samples (using probeset 1419943_s_at for *Ccnb1* as a seed). Our approach identified a ranked list of 1435 probesets, corresponding to 1116 genes, and, after threshold imposition, 259 high-confidence probesets corresponding to 190 unique genes (Figure 3A). Orthologous gene pairs in humans and mice were identified using Affymetrix annotations and the eukaryotic orthologues database InParanoid (O'Brien *et al.*, 2005; Sonnhammer and Ostlund, 2015). Of 195 human genes, 193 had mouse orthologues, and of 190 mouse genes 188 had human orthologues. Intersection of the human and mouse Boolean equivalence list using *CCNB1*/*Ccnb1* as the seeds identified 118 genes in common (Figure 3B: cytokinesis [$n = 8$], centrosome [$n = 4$], kinetochore [$n = 24$], spindle [$n = 6$], chromosome [$n = 15$], cell cycle regulators [$n = 20$], replication and repair [$n = 25$], nuclear pore/periphery [$n = 3$], nucleotide synthesis [$n = 4$], transcription factors [$n = 2$], and other [$n = 7$]). Scatterplots in Figure 3C show strong positive correlation, and logical equivalence between *CDC2* (mouse orthologue *Cdc2a*) *ATAD2*, *CDCA7*, and *CCNB1*,– *CDC2* encodes Cdk1 and was expected based on its central role in cell cycle control. *CDCA7*, mutated in human immunodeficiency-centromeric instability-facial anomalies syndrome, and *ATAD2*, a AAA+ enzyme considered to be a putative oncogene, are both chromatin factors with poorly studied roles in the cell cycle.

The above analysis shows that ~40% of the high-confidence cell cycle genes are not shared between human and mouse.

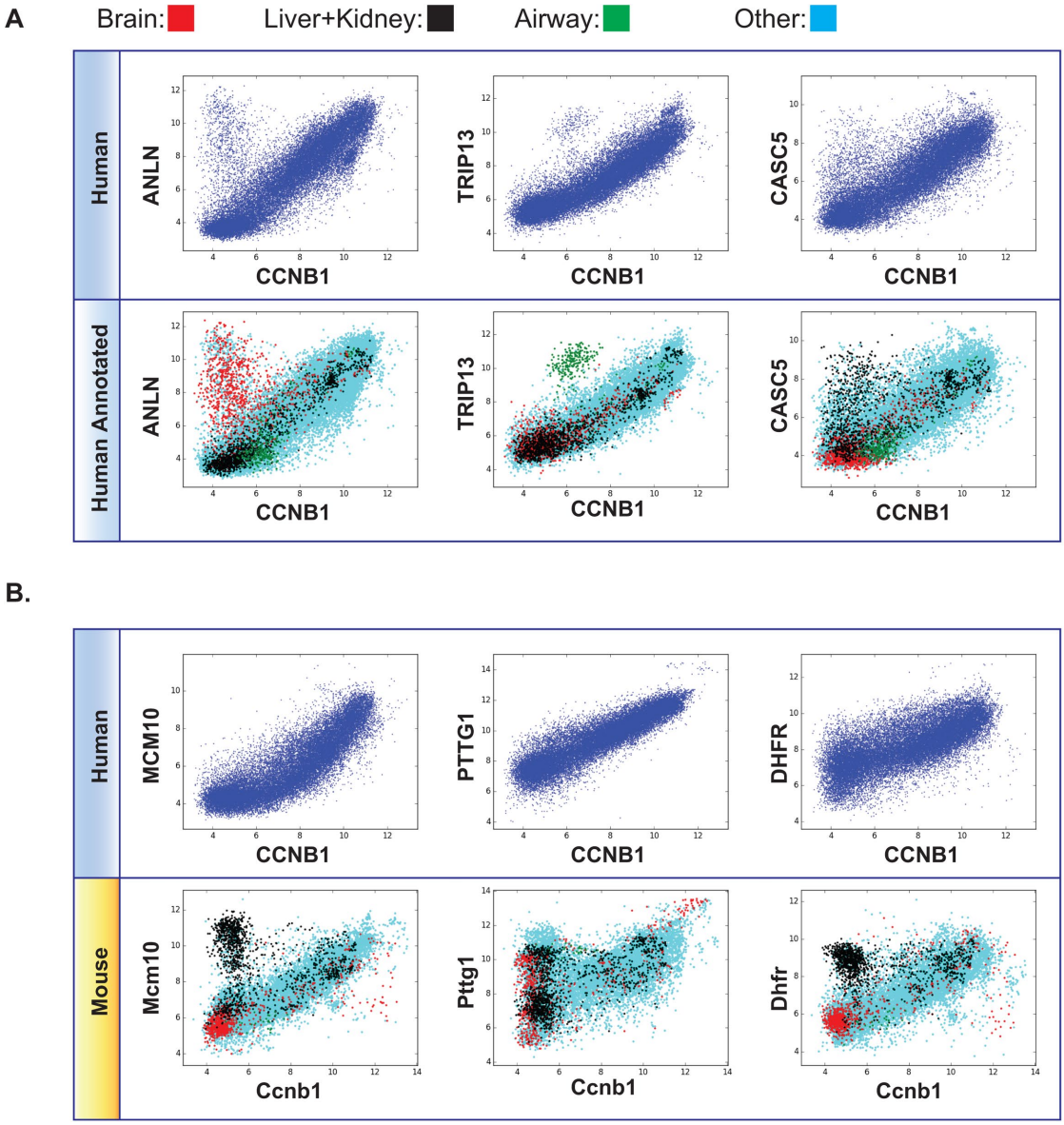

FIGURE 4: Noncell cycle contexts of specific cell cycle genes. (A) Top, Tissue-specific gene expression patterns of three well-known cell cycle genes: *ANLN*, *TRIP13* and *CASC5*. Bottom, Tissue annotation for each sample is highlighted in the scatterplots: Brain (red), Liver+Kidney (black), airway epithelial cells (green), other tissue (blue). In *ANLN* vs. *CCNB1* scatterplot, the outliers are mostly brain samples. In *TRIP13* vs. *CCNB1* scatterplot, the outliers are mostly airway epithelial cells. In *CASC5* vs. *CCNB1* scatterplot, the off-diagonal points are enriched in liver and kidney samples. (B) Top, Three previously known cell cycle genes (*MCM10*, *PTTG1*, and *DHFR*) are equivalent to *CCNB1* in publicly available microarray data in 25,955 human samples. Bottom, Scatterplots between three previously known cell cycle genes (*Mcm10*, *Pttg1*, and *Dhfr*) and *CCNB1* in publicly available microarray data in 11,758 mouse samples. All three scatterplots shows off-diagonal points that are enriched in liver and kidney samples. This demonstrates tissue-specific expression patterns for known cell cycle genes.

However, when human genes were compared with the ranked list of genes before the threshold step in the mouse data set, we observed that 89% (171 genes) were common between human and mouse. This example highlights the value of considering not only the gene lists after thresholding but also the larger ranked list generated after two rounds of BECC. We analyzed whether the remaining 11% difference was due to technical errors or suggestive of species-specific differences. We did not observe probe quality issues, with the dynamic range of gene expression values being high for at least one probeset. Therefore, it is possible that these are real differences, but experimental efforts in mouse and human cells are necessary to test whether this is indeed the case.

Prediction of new contexts for the action of cell cycle genes

The BECC algorithm generates plots of pairwise gene expression analysis across thousands of samples derived from many different sources. Inspection of these plots has the potential to reveal unexpected patterns that may represent potential new, context-specific functions for cell cycle genes. As shown in Figure 4A, *ANLN*, which is well known for its function in cytokinesis, also exhibits a strong off-axis group of points. Similarly, the plot for the gene encoding the AAA+

enzyme *TRIP13*, well studied for its function in the spindle checkpoint and in meiotic chromosome dynamics (Vader, 2015; Corbett, 2017), shows a cluster of points that are not correlated with *CCNB1*. CASC5 is also expressed in a group of samples that have low levels of CCNB1. To assess whether these new contexts represent tissue-specific functions unrelated to cell cycle progression, we developed a dynamic interface, which identifies the data set for each point based on placement of a cursor. Guided by this manual browsing, we highlighted the scatterplots to mark samples based on tissue origins, identified by automated searching of Gene Expression Omnibus (GEO) descriptions of the primary microarray data sets (specifically, by searching for the words brain, liver, hepatocellular, kidney, and airway). As shown in Figure 4A, bottom panel, *ANLN* expression is correlated with that of *CCNB1*, except in brain samples, *TRIP13* shows clear enrichment not correlated with *CCNB1* in airway tissue, and CASC5 shows deviation in liver and kidney samples. For *ANLN*, there is prior work highlighting functions in the nervous system (Tian *et al.*, 2015), whereas for *TRIP13* there is no prior analysis suggesting a function in airway epithelium. While addressing potential new roles for such components will require tissue-specific inhibitions (and potentially strategies that inhibit function in differentiated nondividing cells), these patterns provide the basis to motivate such analysis in the future.

The same approach was employed to identify species-specific differences; for example, the genes encoding the replication factor *MCM10* (*Mcm10*), securin (*PTTG1*, *Pttg1*), and *DHFR* (*Dhfr*) show off-axis expression in liver and kidney samples in the mouse but not in the human data sets (Figure 4B). Such differences may complicate interpretation of genetic perturbations in mice, as observed phenotypes may include consequences of perturbing function in tissue contexts that are not represented in humans.

Expression of cell cycle genes across species: human, mouse, fly, and plants

To investigate expression patterns of cell cycle genes that are shared across species, we performed independent analyses of human, mouse, fly (*Drosophila melanogaster*), and plant (*Arabidopsis thaliana*) data sets. Orthologous gene pairs were identified using Affymetrix annotations and the eukaryotic orthologues database InParanoid (O'Brien *et al.*, 2005; Sonnhammer and Ostlund, 2015). The analyses of human and mouse data sets are described above. Because of the limited number of samples and lower diversity, the fly data set analysis identified slightly higher numbers of probesets and genes: 494 probesets in ProbeSet A, 3295 probesets in ProbeSet B, 790 high scoring probesets, and 771 unique genes. The total number of 77 human probesets and 59 unique genes was shared in human, mouse, and fly analyses. Analysis of the plant data set resulted in 68 probesets in ProbeSet A, 703 probesets in ProbeSet B, 140 high scoring probesets, and 134 unique genes. The total number of 25 human probesets and 19 unique genes was found to be shared in all four species. While the absence of genes may reflect technical issues, these species-specific data sets provide an accessible resource for potential functional follow-up work in these different models.

A resource for analyzing expression of cell cycle genes across species

To make the results of the analysis widely accessible, we developed a user-friendly website resource that enables browsing all of the pairwise expression data (including the dynamic plot browsing interface) and for generating any desired pairwise plot between the set of genes. A detailed set of instructions on how to use the data is at http://hegemon.ucsd.edu/CellCycle (Figure 5). The website resource allows the user to perform following queries.

The website resource provides a set of links to explore the results of the BECC algorithm on human, mouse, fly, and plant data sets. The user can retrieve top N genes by specifying N in a text box and using the topGenes button. The user can then select a gene to plot against a reference gene which is *CCNB1* by default. However, the reference gene can be changed using the makeRef button on the web page. A scatterplot is generated using the reference and

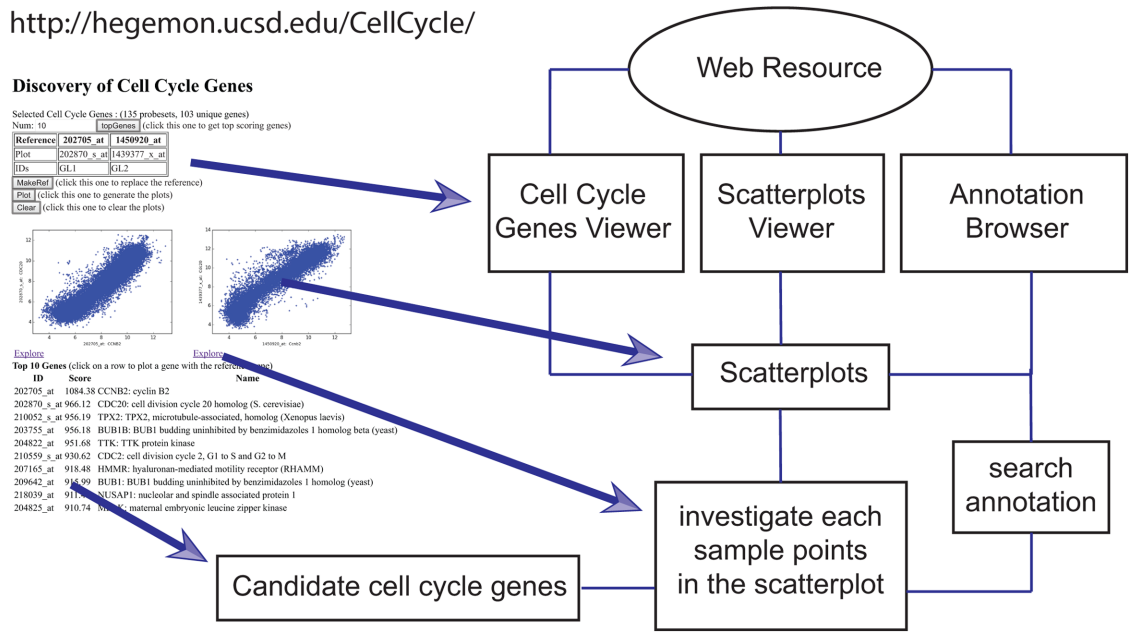

FIGURE 5: Resource. Website to explore the candidate cell cycle genes using three different web interfaces: cell cycle genes viewer, scatterplots viewer, annotation browser. All software packages developed for the analysis are also freely available on the website.

selected probeset IDs using the button Plot. The link that starts with the name conserved provides similar tools to show plots side by side for orthologous pairs across species.

The website resource provides a link where the user can explore scatterplots between any gene-pairs of interest in human, mouse, *Drosophila*, and *Arabidopsis*. All scatterplots are associated with an Explore link that is used to explore the scatterplot in detail. The Explore button on the website shows the scatterplot where the points can be selected by dragging a rectangular area. The selected points appear on the right side as a group with the number of samples. A checkbox next to the group is used to highlight the samples in the scatterplot with a different color. The scatterplot explorer web page provides a button Show which gives the GEO links to the selected experiments.

One of the links in the website resource is dedicated to searching simple words in the GEO annotations in the context of a scatterplot. Using this tool, one can search the tissue types or other metadata that are provided on the GEO annotations web page of the National Center for Biotechnology Information (NCBI) in the context of a scatterplot. For example, it shows the enriched keywords that are present in a set of experiments selected from the scatterplot by mouse click using rectangular areas. Users can search for data related to brain tissues and the specific experiments are highlighted in the scatterplot using a different color. The steps are click global → click submit → select dataset, type gene names, click getPlots → select plots → click explore → select two groups of experiments by mouse click → select groups → click Annotations. To search brain tissue, enter the keyword brain in the textbox next to the Search button after the explore options and click on the Search button. Select the group created to highlight the experiments in the scatterplot.

All software tools and GEO accession numbers used and described above are provided in a link in the website resource. The relevant GEO accession numbers are GSE119083 (*Arabidopsis thaliana*) (Pandey and Sahoo, 2019), GSE119084 (*Drosophila melanogaster*), GSE119085 (mouse), GSE119087 (human), and GSE119128 (collections).

DISCUSSION

Many genes have been implicated in the fundamental cell cycle processes that are critical for proliferation. Cell cycle genes are essential for development and defects in their function or expression are associated with human diseases, such as cancer. Here, we describe an unbiased Boolean approach that identifies cell cycle genes using publicly available gene expression data. We provide ranked lists and high-confidence gene sets after imposition of a threshold in a website resource for human, mouse, *Drosophila melanogaster*, and *Arabidopsis thaliana* (web link). Thirty-eight of the 195 high-confidence human cell cycle genes identified by this approach were not identified by prior synchronization-release expression studies in cultured cells. The difference between our study and previous studies may be due to the types of samples analyzed and the methodology employed in the array-based expression analysis. Notably, 28 of these 38 genes were validated using a new publicly available single-cell RNASeq data set (GSE121265). Only 22 genes from the 195 human genes behaved in a significantly different manner in the mouse data set, consistent with the expectation that the majority of human cell cycle genes should have orthologous mouse cell cycle genes.

A limitation of our approach is that it is dependent on the quality of gene expression measurements. In Affymetrix microarray data sets, there are many probesets for each gene and these can exhibit dramatically different patterns. This variation raises the question of how to choose a representative probeset for each gene. For *CCNB1*, both probesets are similar, making the analysis robust. However, for other genes, we chose the probeset with the best dynamic range and strong signal. We computed the percentage of probesets identified as being a cell cycle gene from the overall probesets for that gene. The percentage for human was 65% and for mouse it was 62%. For specific well-known cell cycle genes, such as E2F1, a good probeset is not available in the Affymetix microarray data (Supplemental Figure S6A); however, E2F1 expression data in TCGA breast cancer RNASeq data are robust and highly correlated with *CCNB1*. Another well-known cell cycle gene *SLBP* has good probesets available in the microarray data sets; however, only the mouse data set exhibited good correlation with *CCNB1* (ranked #307). However, *SLBP* is poorly correlated with *CCNB1* in the human microarray and breast cancer RNASeq data sets and therefore not ranked, suggesting that *SLBP* expression may not be strictly cell cycle-correlated in tissue contexts.

We show here that despite a negative correlation between *CCNB1* and *CCNE2* in cell synchronization experiments, they are positively correlated in analysis of bulk tissue samples. Since the mRNA measurements in bulk tissue represent average gene expression from all cells, expression of *CCNB1* and *CCNE2* appears to be directly proportional to the fraction of cells dividing in tissues, resulting in strong positive correlation in bulk data sets. By focusing on these highly correlated clusters of genes in diverse big data sets, we hope to enrich genes whose function is limited to proliferation in diverse tissues, conditions, and diseases.

Our analysis suggests that certain cell cycles genes, such as *ANLN*, *TRIP13*, and *CASC5*, may function in contexts other than cycling cells (Figure 4A). These genes should be placed in a group of cell cycle genes that have off-axis points; however, it is difficult to assess the percentage of such genes using our approach. For example, while there are many human samples with *CCNB1* low and *ANLN* high (Supplemental Figure S6C), the off-axis points in the human data set are small enough in number to not impact the BooleanNet thresholds. However, in the mouse data set there were sufficient *Ccnb1* low and *Anln* high off-axis points to prevent *Anln* from being Boolean equivalent with *Ccnb1*. Thus, *ANLN* was detected as a cell cycle gene in the human data set but not in the mouse data set. This discrepancy is likely due to sample biases between these data sets and the Boolean approach severely punishing genes that have off-axis points. A modified computational approach will be necessary to classify genes that are similar to these three examples. We note that the scatterplots in our website resource can be queried to assess the identity of samples with off-axis points, enabling users with expertise in specific genes to assess whether there are sufficient such points to motivate experimental efforts.

Extending Boolean equivalence analysis beyond the cell cycle

The approach we describe here is focused on cell cycle genes but is straightforward to extend to any process using a well-chosen seed gene. In addition, the analysis we present is based on microarray data sets but is directly applicable to RNA-Seq data sets. The normalization steps involved are distinct for the two types of expression data (Zhao *et al.*, 2014). To facilitate analysis on a different process employing a new seed gene, we have deposited for open access the normalized human microarray data (at GSE119087) and the required software (at http://hegemon.ucsd.edu/CellCycle/Software/). A detailed set of instructions on how to conduct such an analysis is presented in the software links on the main website resources. Notably, the analysis requires storage space (50 GB) and computing power that are present in typical current desktop/laptop computers, making this approach accessible to all.

MATERIALS AND METHODS

Data collection and annotation

Publicly available microarray databases in Human U133A (n = 21,285, GPL96), Human U133 Plus 2.0 (n = 25,955, GPL570), Mouse 430 2.0 (n = 11,758, GPL1261), Affymetrix Drosophila Genome 2.0 Array (n = 2,687, GPL1322), and *Arabidopsis thaliana* ATH1 (n = 4,306, GPL198) Affymetrix platform (Pandey and Sahoo, 2019) were downloaded from the NCBI GEO website (Edgar *et al.*, 2002). Gene expression summarization was performed by normalizing each Affymetrix platform by Robust Multichip Average (Irizarry *et al.*, 2003). A single-cell RNASeq data set that quantifies continuous cell cycle phase using single-cell gene expression data (GSE121265) was used for validation (Hsiao *et al.*, 2019). We considered all gene names annotated at NCBI (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene_info.gz, downloaded on July 12, 2018) in our comparisons with prior work.

Boolean analysis of data sets

The expression values of each gene were ordered from low to high and a rising step function was computed to define a threshold by StepMiner algorithm in the individual data set (Sahoo *et al.*, 2007). If the assigned threshold for a gene was t, then expression levels above $t + 0.5$ were classified as high, and the expression levels below $t - 0.5$ were classified as low. Expression levels between $t - 0.5$ and $t + 0.5$ were classified as intermediate. Previously published BooleanNet algorithm was performed to determine Boolean Implication relationships between genes (Sahoo *et al.*, 2008). Briefly, BooleanNet algorithm searches for at least one sparsely populated quadrant in a scatterplot between two genes. The intermediate expression values were ignored by the BooleanNet algorithm. There were six possible scenarios: one of the four quadrants was sparse (four asymmetric Boolean implications) and two diagonally opposite quadrants were sparse (Equivalent and Opposite Boolean implications).

BECC analysis

BECC analysis is based on Boolean equivalent relationships, pairwise correlation, and linear regression analysis (Supplemental Figure S1). A gene pair was included in the BECC analysis if they had a Boolean equivalent relationship or both had a Boolean equivalent relationship with a common third gene. This analysis was performed in two steps. First, a selected probeset of a seed gene was used as a starting point to identify a list of probesets (ProbeSet A) that are Boolean equivalent to the selected probeset. Next, this list was expanded (ProbeSet B, L) by identifying other probesets that are Boolean equivalent to at least one of the probeset from ProbeSet A. A score was computed for a pair of probesets from L by using the correlation r and slope of fitted line s (if $s > 1$, $1/s$ was used as slope):

$$\text{score} = r^2 + s^2$$

The score is a number between 0 and 2 given $r > 0$ and $s > 0$. A matrix of scores M was computed for all probesets in L. Every row of this matrix was sorted based on the score in ascending order. The whole matrix was then multiplied using a column vector of ranks: *[0 1 2 … len(L)-1]*. In other words, the score for the probeset in row i gs_i was computed as follows:

$$gs_i = \frac{1}{len(L)} \sum_{k=0}^{len(L)-1} k * \text{score}_{ik} / 2$$

where score_{ik} is the k*th* smallest score for the probeset in row i.

The StepMiner algorithm was used to compute a threshold to identify the high-scoring probesets gs_i. The final result of the BECC is this list of high-scoring probesets.

Statistical justification

Empirical distribution of the pairwise gene scores was computed for each of our data sets by randomly selecting pairs of probesets (Supplemental Figure S3). Using this distribution, average probeset score $E[gs_i]$ and stddev(gsi) can be estimated as follows:

$$E[gs_i] = \frac{1}{len(L)} \sum_{k=0}^{len(L)-1} k * \frac{E[\text{score}_{ik}]}{2} = E[\text{score}] * \frac{len(L)-1}{2}$$

$$\text{stddev}(gs_i) = \sqrt{\text{Variance}[\text{score}] * \frac{len(L)-1}{2}}$$

The p value for the StepMiner identified threshold was computed using a Z-test. All statistical tests were performed using statistical programming language R version 3.2.3 (2015-12-10).

Data submission

All the data generated in the described analyses are submitted to GEO: GSE119083 (*Arabidopsis thaliana*) (Pandey and Sahoo, 2019), GSE119084 (*Drosophila melanogaster*), GSE119085 (mouse), GSE119087 (human), and GSE119128 (collections).

Data access

http://hegemon.ucsd.edu/CellCycle/:

GSE119083—Plant Boolean Implication Network.
GSE119084—Fly Boolean Implication Network.
GSE119085—Mouse Boolean Implication Network.
GSE119087—Human Boolean Implication Network.
GSE119128—An unbiased Boolean analysis of public gene expression data for cell cycle gene classification.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grants #R00-CA151673 to D.S. and GM074215 to A.D., 2017 Padres Pedal the Cause/Rady Children's Hospital Translational PEDIATRIC Cancer Research Award (Padres Pedal the Cause/RADY #PTC2017) to D.S., and 2017 Padres Pedal the Cause/C3 Collaborative Translational Cancer Research Award (San Diego NCI Cancer Centers Council [C3] #PTC2017) to D.S. We thank Beata Mierzwa for her critical review of the draft and giving us feedback.

REFERENCES

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25, 25–29.

Bar-Joseph Z, Siegfried Z, Brandeis M, Brors B, Lu Y, Eils R, Dynlacht BD, Simon I (2008). Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells. Proc Natl Acad Sci USA 105, 955–960.

Bertoli C, Klier S, McGowan C, Wittenberg C, de Bruin RA (2013). Chk1 inhibits E2F6 repressor function in response to replication stress to maintain cell-cycle transcription. Curr Biol 23, 1629–1637.

Bertomeu T, Coulombe-Huntington J, Chatr-Aryamontri A, Bourdages KG, Coyaud E, Raught B, Xia Y, Tyers M (2018). A high-resolution genome-wide CRISPR/Cas9 viability screen reveals structural features and contextual diversity of the human cell-essential proteome. Mol Cell Biol 38, e00302-17.

Blomen VA, Majek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, Staring J, Sacco R, van Diemen FR, Olk N, Stukalov A, *et al.* (2015). Gene essentiality and synthetic lethality in haploid human cells. Science 350, 1092–1096.

Brandeis M, Rosewell I, Carrington M, Crompton T, Jacobs MA, Kirk J, Gannon J, Hunt T (1998). Cyclin B2-null mice develop normally and are fertile whereas cyclin B1-null mice die in utero. Proc Natl Acad Sci USA 95, 4344–4349.

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, *et al*. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 2, 65–73.

Corbett KD (2017). Molecular mechanisms of spindle assembly checkpoint activation and silencing. Prog Mol Subcell Biol 56, 429–455.

DeBruhl H, Wen H, Lipsick JS (2013). The complex containing Drosophila Myb and RB/E2F2 regulates cytokinesis in a histone H2Av-dependent manner. Mol Cell Biol 33, 1809–1818.

de Lichtenberg U, Jensen LJ, Fausboll A, Jensen TS, Bork P, Brunak S (2005). Comparison of computational methods for the identification of cell cycle-regulated genes. Bioinformatics 21, 1164–1171.

Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003). DAVID: Database for annotation, visualization, and integrated discovery. Genome Biol 4, P3.

Edgar R, Domrachev M, Lash AE (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30, 207–210.

Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, Haw R, Jassal B, Korninger F, May B, *et al*. (2018). The Reactome Pathway Knowledgebase. Nucleic Acids Res 46, D649–D655.

Fischer M, Grossmann P, Padi M, DeCaprio JA (2016). Integration of TP53, DREAM, MMB-FOXM1 and RB-E2F target gene analyses identifies cell cycle gene regulatory networks. Nucleic Acids Res 44, 6070–6086.

Gauthier NP, Jensen LJ, Wernersson R, Brunak S, Jensen TS (2010). Cyclebase.org: version 2.0, an updated comprehensive, multi-species repository of cell cycle experiments and derived analysis results. Nucleic Acids Res 38, D699–D702.

Gauthier NP, Larsen ME, Wernersson R, de Lichtenberg U, Jensen LJ, Brunak S, Jensen TS (2008). Cyclebase.org—a comprehensive multi-organism online database of cell-cycle experiments. Nucleic Acids Res 36, D854–D859.

Giotti B, Joshi A, Freeman TC (2017). Meta-analysis reveals conserved cell cycle transcriptional network across multiple human cell types. BMC Genomics 18, 30.

Grant GD, Brooks L 3rd, Zhang X, Mahoney JM, Martyanov V, Wood TA, Sherlock G, Cheng C, Whitfield ML (2013). Identification of cell cycle-regulated genes periodically expressed in U2OS cells and their regulation by FOXM1 and E2F transcription factors. Mol Biol Cell 24, 3634–3650.

Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, Mis M, Zimmermann M, Fradet-Turcotte A, Sun S, *et al*. (2015). High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. Cell 163, 1515–1526.

Hosack DA, Dennis G Jr, Sherman BT, Lane HC, Lempicki RA (2003). Identifying biological themes within lists of genes with EASE. Genome Biol 4, R70.

Hsiao CJ, Tung P, Blischak JD, Burnett J, Barr K, Dey KK, Stephens M, Gilad Y (2019). Characterizing and inferring quantitative cell cycle phase in single-cell RNA-seq data analysis. bioRxiv, 526848.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP (2003). Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31, e15.

Lewis PW, Sahoo D, Geng C, Bell M, Lipsick JS, Botchan MR (2012). Drosophila lin-52 acts in opposition to repressive components of the Myb-MuvB/dREAM complex. Mol Cell Biol 32, 3218–3227.

Marguerat S, Jensen TS, de Lichtenberg U, Wilhelm BT, Jensen LJ, Bahler J (2006). The more the merrier: comparative analysis of microarray studies on cell cycle-regulated genes in fission yeast. Yeast 23, 261–277.

Morgan DO (2006). The Cell Cycle: Principles of Control, London, UK: New Science Press.

O'Brien KP, Remm M, Sonnhammer EL (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Res 33, D476–D480.

Pandey S, Sahoo D (2019). Identification of gene expression logical invariants in Arabidopsis. Plant Direct 3, e00123.

Peña-Diaz J, Hegre SA, Anderssen E, Aas PA, Mjelle R, Gilfillan GD, Lyle R, Drablos F, Krokan HE, Saetrom P (2013). Transcription profiling during the cell cycle shows that a subset of Polycomb-targeted genes is upregulated during DNA replication. Nucleic Acids Res 41, 2846–2856.

Sadasivam S, DeCaprio JA (2013). The DREAM complex: master coordinator of cell cycle-dependent gene expression. Nat Rev Cancer 13, 585–595.

Sahoo D, Dill DL, Gentles AJ, Tibshirani R, Plevritis SK (2008). Boolean implication networks derived from large scale, whole genome microarray datasets. Genome Biol 9, R157.

Sahoo D, Dill DL, Tibshirani R, Plevritis SK (2007). Extracting binary signals from microarray time-course data. Nucleic Acids Res 35, 3705–3712.

Santos A, Wernersson R, Jensen LJ (2015). Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. Nucleic Acids Res 43, D1140–D1144.

Sim CK, Perry S, Tharadra SK, Lipsick JS, Ray A (2012). Epigenetic regulation of olfactory receptor gene expression by the Myb-MuvB/dREAM complex. Genes Dev 26, 2483–2498.

Sonnhammer EL, Ostlund G (2015). InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. Nucleic Acids Res 43, D234–D239.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, *et al*. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 102, 15545–15550.

Tian D, Diao M, Jiang Y, Sun L, Zhang Y, Chen Z, Huang S, Ou G (2015). Anillin regulates neuronal migration and neurite growth by linking RhoG to the actin cytoskeleton. Curr Biol 25, 1135–1145.

Vader G (2015). Pch2(TRIP13): controlling cell division through regulation of HORMA domains. Chromosoma 124, 333–339.

Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM (2015). Identification and characterization of essential genes in the human genome. Science 350, 1096–1101.

Wang BH, Lim JW, Lim JS (2016). Gene regulatory network identification from the yeast cell cycle based on a neuro-fuzzy system. Genet Mol Res 15.

Wen H, Andrejka L, Ashton J, Karess R, Lipsick JS (2008). Epigenetic regulation of gene expression by Drosophila Myb and E2F2-RBF via the Myb-MuvB/dREAM complex. Genes Dev 22, 601–614.

Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO, *et al*. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol Biol Cell 13, 1977–2000.

Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One 9, e78644.