Modeling bi-modality improves characterization of cell cycle on gene expression in single cells

Andrew McDavid[a,b,*], Lucas Dennis[c,*], Patrick Danaher[c], Greg Finak[b], Michael Krouse[c], Alice Wang[d], Philippa Webster[c], Joseph Beechem[c], Raphael Gottardo[a,b,1]

[*] Authors contributed equally to this work

Affiliations

[a] Department of Statistics, University of Washington, Seattle, WA 98195, USA

[b] Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

[c] NanoString Technologies, Seattle, WA 98118, USA

[d] BD Biosciences, San Jose CA 95131, USA

[1] To whom correspondence should be addressed.

Abstract

Advances in high-throughput, single cell gene expression are allowing interrogation of cell heterogeneity. However, there is concern that cell cycle might bias characterizations of gene expression at the single-cell level. We assess the effect of cell cycle phase on gene expression in single cells by measuring 333 genes in 930 cells across three phases and three cell lines. We determine each cell's phase non-invasively without chemical arrest, and use it as a covariate in tests of differential expression. We observe bi-modal gene expression, a previously-described phenomenon, wherein the expression of otherwise abundant genes is either strongly positive; or undetectable within individual cells. This bi-modality is likely both biologically and technically driven. Irrespective of its source, we show that it should be modeled to draw accurate inferences from single cell expression experiments. To this end, we propose a semi-continuous modeling framework based on the generalized linear model, and use it to characterize genes with consistent cell cycle effects across three cell lines. Our new computational framework improves the detection of previously characterized cell-cycle genes compared to approaches that do not account for the bi-modality of single-cell data. We use our semi-continuous modelling framework to estimate single cell gene co-expression networks. These networks suggest that in addition to having phase-dependent shifts in expression (when averaged over many cells), some, but not all, canonical cell cycle genes tend to be co-expressed in groups in single cells. We estimate the amount of single cell expression variability attributable to the cell cycle. We find that the cell cycle explains only 5%-17% of expression variability, suggesting that the cell cycle will not tend to be a large nuisance factor in analysis of the single cell transcriptome.

Introduction

With the advent of single cell expression profiling [1–4], the assessment of cell population heterogeneity and identification of cell subpopulations from mRNA expression is achievable [5–7]. However, at the single cell level, there is concern that cell cycle might interfere with the characterization of gene expression variability [8]. As many biological samples are prepared from asynchronous cell populations, where each cell is in an unknown phase of the cell cycle, it is imperative to understand the impact of cell cycle in order to account for its effect on observed expression patterns and downstream data analysis. Here, we have measured mRNA expression and cell cycle from 930 single cells derived from three cell lines in order to explore this hypothesis.

A distinctive feature of single-cell gene expression data is the bimodality of expression values. Genes can be on (and a positive expression measure is recorded) or off (and the recorded expression is zero or negligible)[9,10]. This dichotomous characteristic of the data prevents use of the typical tools of designed experiments such as linear modeling and analysis of variance (ANOVA). We develop a novel computational framework to overcome this problem. First, a probabilistic mixture model-based framework allows the separation of positive expression values from background noise using gene-specific thresholds. After signal separation by thresholding, we model separately the frequency of expression (the fraction of cells expressing a gene) and the continuous, positive expression values. Our semi-continuous framework combines evidence from the two salient parameters of single cell expression in a statistically appropriate manner, an approach dubbed the Hurdle model[11,12]. Our new framework allows for testing arbitrary contrasts and allows the use of variance components/mixed models, thus bringing to bear the full power of the general linear model.

The Hurdle model allows us to identify many genes with an archetypal cell cycle expression pattern despite a frequently bimodal distribution of expression. It also suggests that stochastic variation in single cell gene expression is relatively large compared to the effect of cell cycle. We find that even in the most tightly regulated genes, cell cycle explains only 27% of the variability, while in the median gene in our data set, cell cycle explains 5%-17% of the variability. The semi-continuous model also provides a framework for estimating co-expression networks – in which edges connect genes whose partial correlations remain after removing the effect of all other genes – while adjusting for population-level nuisance factors that could bias network inference. Applying this novel framework to our data, we show that only a subset of canonical cell cycle genes are highly co-expressed in single cells.

Results

Periodic expression associated with cell cycle is observed at the single-cell level

In order to assess differential expression associated with actively cycling cells, expression of 333 genes was interrogated in 930 cells, across three cell lines: H9 (HTB-176), MDA-MB-231 (HTB-26), and PC3 (CRL-1435) (Figure 1A). Single cell expression was measured from flow-sorted cells across three different cell lines and compared between cell cycle phases and cell lines via nCounter single cell profiling, a multiplexed hybridization-based detection technology that utilizes fluorescent barcodes to count individual target nucleic acid molecules [13]. This platform has been recently adapted to enable expression profiling from

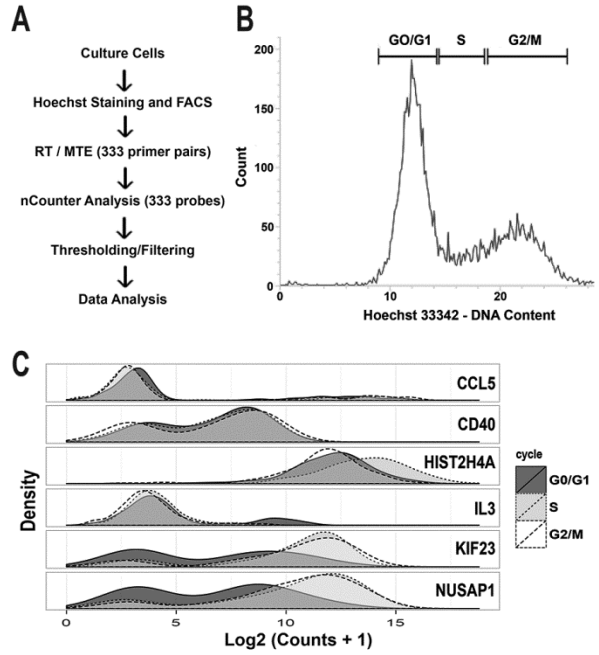

Figure 1. **Individual cells were flow sorted by DNA content, and gene expression profiled.** (A) H9, MB231, and PC3 cells were cultured and sorted into lysis buffer. The resulting lysate was amplified via multiplexed target enrichment (MTE) and digital counts of expression were optically read via nCounter. (B) Individual cells were sorted into three populations based on retention of Hoechst dye (G0/G1, S and G2/M). (C) The density distribution of log counts for each gene was generally bimodal with some genes showing clear changes in distribution between cell cycle phase.

single cells via hybridization after a multiplexed target enrichment (MTE) in which mRNA is first converted to cDNA and then amplified [14].

Each cell was categorized as being in G0/G1, S or G2/M phase by measuring DNA content via flow-cytometry based on retention of Hoechst dye (Figure 1B and S1)[15]. Probes were selected for cell cycle associated genes (n = 119). These genes provided coverage of the entire cell cycle (Data Set S1) based on peak expression and periodicity information obtained from Cyclebase, an integrated database of bulk cell cycle expression profiling experiments that scores and ranks genes based on strength of evidence for a cell cycle associated expression pattern[16]. Probes were also included for non-cell-cycle associated genes with primary roles in the inflammatory response, and housekeeping controls without a Cyclebase ranking (n=214). We denote probes with a Cyclebase rank (i.e. genes with the strongest evidence for cell cycle associated periodic expression) as the *ranked* set.

253 genes were expressed and passed quality control (see Methods). Genes showed a bimodal expression pattern in log-transformed mRNA levels (Figure 2), consistent with a burst-model of "on/off" transcription at the single cell level [17] and consistent with the kinetics of PCR amplification with low starting template concentrations, described by us and other authors [9,10].

Expression levels for each gene were most different between cell lines (Figure 2 and S2). Many genes, including those in the *ranked* set showed cell line-specific expression patterns. For example, expression of TOP2A in G0/G1 varied from 70% of cells in MB-231 and PC3 to nearly universal in H9. This cell line effect was a nuisance factor we needed to adjust for in differential expression tests on cell cycle.

Nonetheless, many genes from the *ranked* set, such as KIF23, TOP2A, HJURP, NUSAP1, and TPX2 exhibited expression patterns consistent with cell cycle regulation (Figure 2). Figure 2 also reveals that changes in both the positive expression mean (i.e. the mean over the cells expressing that gene; PEM), and changes in the frequency of cells expressing a gene, occur throughout the cell cycle. The frequency and PEM in these genes also vary widely between cell lines, so it was important to adjust for cell

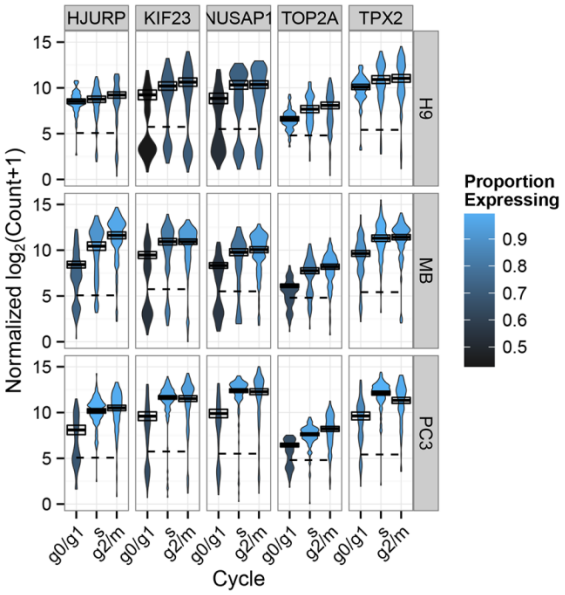

Figure 2. Top 5 cell cycle genes detected by the Hurdle model. A violin plot shows the density of log counts of mRNA in each condition. The expression threshold estimated for each gene is shown as a dashed line, so that the ratio of area above the dashed line reflects the proportion of cells expressing a gene. Blue shades of the violin depict genes with more expressing cells in a condition. The positive mean and 95% confidence interval is depicted as a box with solid line. The Hurdle model (see Methods) combines evidence for changes in either of these parameters, after adjusting for cell line, to determine statistical significance.

line effects for accurate assessment of differential expression (Figure S4).

In order to test for significant differences in expression between cell cycle phases that were consistent across cell lines, we developed an ANOVA-like model (Hurdle model, see Methods) that permits adjustment for additive effects due to cell line. The Hurdle model improves the power to detect changes in single-cell expression by testing both the frequency of expression (corresponding to the relative distribution of cells between the two modes), and the PEM. Combining evidence from the discrete and continuous components of the data provides better sensitivity to changes in expression compared to test statistics based on frequencies of expression (discrete) or on the PEM (continuous) alone (Figure S3 and S4).

Within the three cell lines tested here, *significant* differential expression (Bonferroni-adjusted for 253 tests at P<0.05) was observed for 78 genes in the *ranked* set and 28 genes in the *unranked* set (Figure 3A). Genes showing the strongest cell cycle associated expression patterns in bulk measurements were more likely to be identified as significant in the single-cell populations (Figure 3A-B).

For each gene, peak time was determined based on the phase (G0/G1, S or G2/M) with maximum average expression across all cell lines. Despite large cell-line-specific expression variability, peak times were broadly consistent with Cyclebase annotations (Figure 3C), and especially so within the subset of genes with strongest evidence of cycle regulation in our data (e.g. Bonferroni significant at P<0.05).

The majority of genes in the *unranked* set (115/143 or 80%) did not exhibit significant cell cycle effects, in concordance with their primary roles in functions unrelated to the cell cycle. Of the 28 *unranked* genes that exhibited a significant cell cycle phase association, we noted genes involved in cytoskeletal organization (PLAT), proliferation (PDGFA), and signaling pathways (IFNA1, IFNB1) that have been previously demonstrated to modulate progression through the cell cycle[18].

Cell cycle explains a small portion of the gene expression variability

It has been argued that a substantial portion of the stochastic variability observed in single cell gene expression experiments may be caused

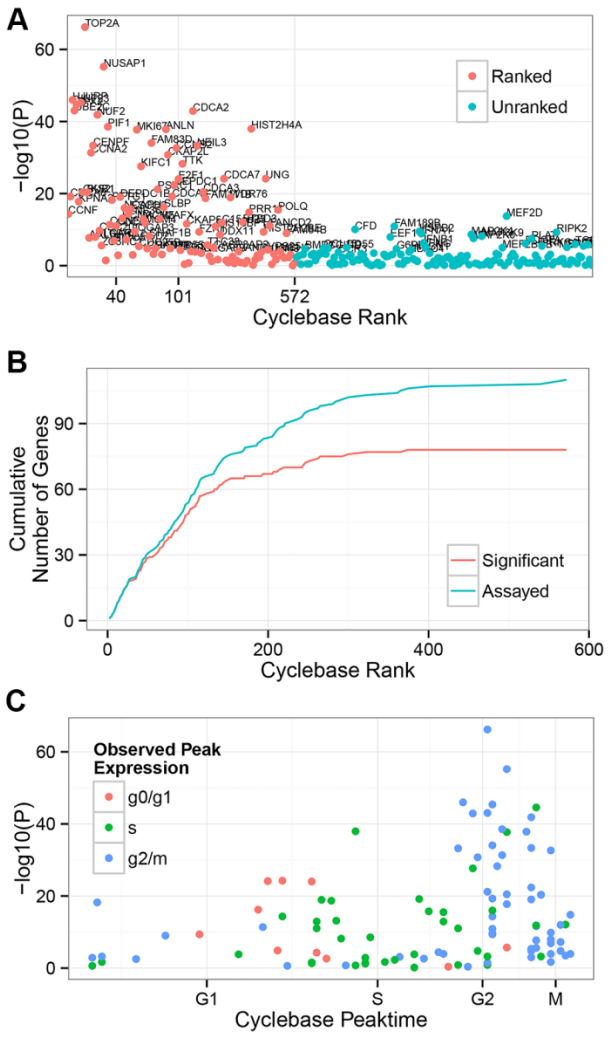

Figure 3. The Hurdle model identifies genes with cell cycle phase-dependent expression. (A) Hurdle model strength of evidence of cell cycle phase dependent expression for genes within our panel versus Cyclebase rank. P-values (-log10(p)) are shown on the y-axis. *Ranked* genes (red) are ordered on the x-axis according to Cyclebase rank. Unranked genes (blue) appear in alphabetical order. Genes significant after Bonferroni adjustment are annotated with their names. (B) Cumulative number of significant (red) or all (blue) *ranked* genes versus Cyclebase rank. Genes with lower Cyclebase rank, and hence stronger evidence of cycle regulation in bulk expression, are detected more often than genes with weaker evidence as shown by the minimal gap between significant (red) and all (blue) *ranked* gene lines at Cyclebase rank < 150. (C) Hurdle model strength of evidence of cell cycle phase dependent expression in *ranked* genes versus phase of peak expression estimated from bulk data in Cyclebase. Experimentally observed peak times broadly match the times estimated from bulk data. Concordance in observed peak times is greater for genes with stronger evidence of differential expression.

by global changes in transcription due to cell cycle [19]. We explore this idea by examining the proportional change in the Hurdle model fit associated with inclusion and omission of cell cycle as an explanatory variable. Because the Hurdle model accounts for both the dichotomous (on/off) and continuous nature of single cell data, the change in deviance (generalized linear model log-likelihood) between nested models can be used calculate the amount of variability explained by cell cycle. The total deviance can be partitioned into components corresponding to cell cycle effects, nuisance effects described below, and residual effects. The ratio of cell cycle deviance to the sum of cell cycle plus residual deviance can then be interpreted as the analog to the coefficient of determination in linear least squares.

We consider expression changes due to main effects and interactions of *cell cycle* by *cell line* and account for amplification efficiency and average cell line effect (see Materials and Methods). Only modest amounts of the single cell expression variability can be explained by cell cycle (Figure 4). Within the *ranked* gene set, cell cycle phase explains 8% of the deviance in the median gene and 27% of the deviance in the top gene (TOP2A). In *unranked* genes, phase explains only 5% of the deviance in the median gene.

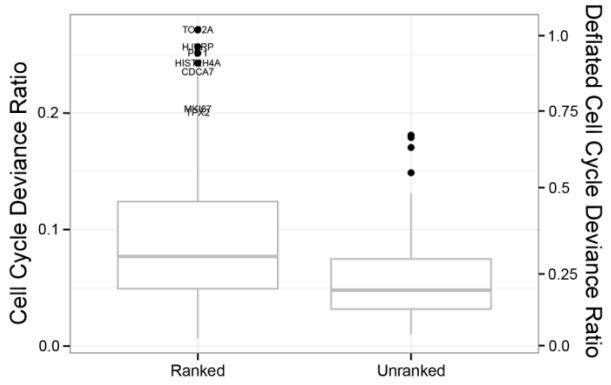

Figure 4. Box and Whiskers Plot of Cell Cycle Deviance Ratio in Ranked and Unranked Genes. The proportion of stochastic variability in the Hurdle Model explained by cell cycle is shown on the primary y-axis (left) for *ranked* and *unranked* genes, with box giving the 25th, 50th and 75th percentiles, and whiskers showing 1.5 times the inter-quartile range. The deflated scale on the secondary y-axis (right) shows the deviance as a percentage of the most completely explained gene (TOP2A, 27%) and is intended as an upper bound for the amount of remaining biological deviance in non-cell-cycle genes. Under this conservative rescaling, cell cycle explains only 25% of the deviance in 75% of unranked genes.

To derive these estimates, it is important to be able to account for the nuisance factors by using the Hurdle model. If cell-to-cell variation in amplification efficiency is not removed, we underestimate the explanatory power of cell cycle on in the median ranked gene by 26% since the unmodeled deviance would include this large additional component. Similarly, other unmeasured factors may inflate the residual deviance and attenuate the apparent role of cell cycle. These factors could include errors in inferring the cell cycle phase via FACS or imperfect modeling of changes in amplification or detection efficiency between samples. To guard against this attenuation, we set an upper bound on cell-cycle-dependent variation as follows: We suppose that transcription of the gene with the most deviance attributable to cell cycle (TOP2A, 27%) would be entirely regulated in a phase-dependent manner, and we characterize other genes' cell-cycle-dependent deviance relative to this maximum. For example, a gene with 13.5% cell-cycle-dependent deviance has half as strong a cell cycle effect as TOP2A, leading to the conclusion that at most 50% of this gene's deviance could be attributable to cell cycle. Even under these generous upper bounds, cell cycle phase explains only 18% (eg, .05/.27) and 29% (eg,.08/.27) of the deviance in the median gene in the *unranked* and *ranked* sets, respectively, suggesting that even when allowing for cell line-specific cell cycle effects, cycle is generally a small factor in gene expression variability in the human transcriptome.

Network analysis reveals gene co-expression at the single-cell level

Single-cell gene expression data sets have the resolution to reveal not only differential expression in response to biological variables like cell cycle phase, but also to provide insight into co-expression between genes at the cellular level (e.g. the influence of one gene on another's expression or the sharing of upstream regulatory elements). In bulk-gene expression data (e.g. microarrays), apparent co-expression arises from tissue-level factors inducing shared marginal changes in genes. For example, different radiation doses in samples will induce correlation amongst all the genes affected by radiation, regardless of whether these genes interact or even participate in the same biological processes. In contrast, single cell data allow us to isolate co-expression arising from cellular-level factors, giving us access to more fundamental biological relationships. If

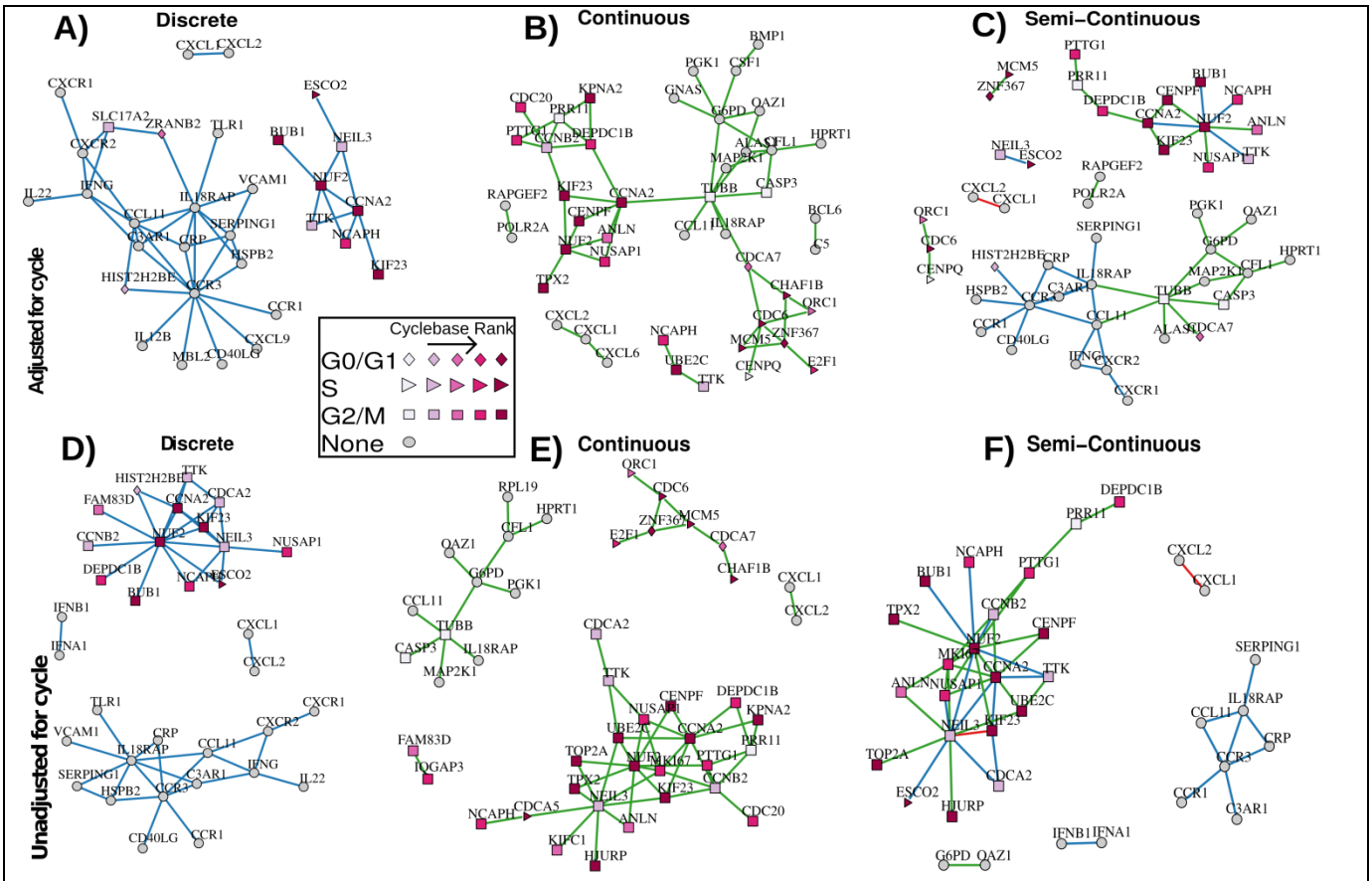

Figure 5. Coexpression networks estimated using the Hurdle Model. Data from three cell lines and three cycles are combined and adjusted for additive effects of cell line and pre-amplification efficiency. Networks of the top 60 edges (ranked by partial correlation) using logistic regressions on discretized expression (A,D), linear regressions on positive, continuous expression values (B,E), and combining the top 30 edges from discrete and continuous components are shown (C,F). Panels A-C adjust for additive cell cycle effects, while panels D-F are unadjusted. The shape of the glyph corresponds to the cycle with peak expression from cyclebase, while the saturation of the glyph corresponds to the ranking. Blue and green edges are partial correlations detected from discrete expression and continuous expression, respectively. Red edges are detected in both discrete and continuous expression.

two genes are correlated across cells drawn from the same environment, then the two genes are likely to share an intimate biological relationship: they may be regulated by the same transcription factor, or one gene may directly regulate the other. The distinction between cellular and marginal co-expression follows from a probabilistic identity on conditional covariances (see Materials and Methods).

When cell cycle is not adjusted for (Figure 5 D-F), known cell cycle genes with strong evidence of marginal regulation comprise the majority of the network. These genes generally peak in phase G2/M, suggesting that the co-expression is mostly driven by the coincident peak in average expression. The networks adjusted for cell cycle remove these marginal effects (Figure 5 A-C). In some cell cycle genes, substantial evidence for co-expression remains, but now additional co-expression is detected in

genes without a previously described cell cycle role. If cell cycle is not conditioned upon, then the strong marginal shifts in expression in canonical cell cycle genes overwhelm subtler co-expression in unranked genes.

Combining both discrete and continuous networks allows a richer set of genes to be characterized. When discrete expression is used alone, networks primarily consist of G2/M peaking genes and unranked genes (Figure 5A). When positive, continuous expression is also used, S and G0/G1 peaking genes enter the networks (Figure 5B-C). The semi-continuous combination of the two networks (with the top 30 edges from discrete and continuous networks) made possible by the Hurdle model leads to sparser networks with smaller average degree.

The adjusted, semi-continuous network depicted in Figure 5C consists of two primary sub-networks, one consisting entirely of ranked genes, and another largely consisting of weakly ranked and unranked genes. The persistence of a subset of ranked genes suggests that this subset is co-expressed at the single cell level as opposed to being co-expressed on average at the population level. The sub-network of ranked genes contains the central node of NUF2, a highly-conserved protein required for stable kinetochore localization of centromere-associated protein E (CENP-E) [20]. NUF2 is connected to other actors in mitotic organization such as ANLN, KIF23, and CENPF, as well as the check-point genes CCNA2 and BUB1, reflecting the central role of these genes in mitosis. The sub-network of primarily unranked genes contains two key nodes: TUBB and CCR3. The predominance of genes associated with cell growth, like TUBB, and transmembrane proteins, like CCR3, in the unranked cluster is likely related to the actively dividing nature of the profiled cells, i.e. dividing cells must generate new scaffolding and membrane-related materials to support growth. This relatively large sub-network of unranked and weakly ranked genes is largely missed by the unadjusted analysis that is biased by the population level cell-cycle effect.

Discussion

Stochastic, bimodal expression is a hallmark of single cell data [21–23]. Within a population of cells, detectable expression for any given gene typically resides in one of two modes, corresponding to an “on” or “off” state. Both technical and biological factors likely contribute to this bimodality. Quantities of some species of cDNA may be minute after reverse-transcription, and in this case random variation in the number of template-primer-enzyme complexes that form during each annealing phase may dominate the kinetics of the PCR [24]. But regardless of its origin, modeling bimodality improves the power of differential expression tests.

By accounting for these features of single cell data while allowing for complex study designs, the Hurdle model is an important step forward for single-cell analysis. We demonstrate the model’s ability to identify many genes with a periodic expression pattern from asynchronously cultured cells utilizing a combination of FACS sorting and these new analytical techniques, including genes with little previous evidence of cell cycle associated periodic expression like MEF2D [25] and FAM189B. The Hurdle model is able to identify phase-dependent patterns of expression despite the fact that G2 and M phases are indistinguishable by DNA content. The similar rank ordering of differentially expressed genes in our single cell experiment as compared to bulk experiments and concordance in the phase of peak expression demonstrates the power of the Hurdle model. While we have applied the Hurdle model to our specific problem, the

approach is general and can be applied to test any effect of interest in a single-cell gene expression dataset. We offer this modeling framework as an R package for other interested users at github.com/RGLab/SingleCellAssay.

Single cell data also allows unparalleled resolution of genes' co-expression patterns. While bulk expression data can reveal correlation induced by varying biological conditions, single-cell data has the possibility to reveal co-expression driven by shared regulatory elements within the cell. However, when inferring gene expression networks, it is important to adjust for population level covariates that could bias the network estimation, especially for genes that are marginally affected by such a population level covariate (like known cell cycle genes in our experiment.) By measuring a limited set of cell cycle associated genes, we are able to identify a network of co-expressed genes with known roles in cell cycle regulation even after adjusting for cell cycle phase.

It is crucial to understand the relationship between cell cycle and the stochastic nature of single cell expression as it determines the magnitude of the cell cycle's distorting effect on single cell analyses. In contrast to earlier estimates of Zopf *et al*. [19] we find little evidence of periodic regulation of expression among non-cell cycle associated genes. Our results are consistent with genome-wide mRNA profiling efforts utilizing bulk expression methodologies in mammalian cells where genes with cycle-dependent periodic expression patterns are limited and well-characterized [16,26,27]. Disparity between our findings and those of Zopf *et al*. may arise from differences between yeast and mammalian cells. Moreover, Zopf *et al.* primarily focus on a single, synthetic promoter while we sample hundreds of transcripts presumably driven by many different promoters. Whether the substantial remaining variability is inherent to the human single cell, or due to thus far latent, unmeasured biological variables remains to be explored.

Materials and Methods

Cell Lines and Flow Cytometry

Three human cell lines H9 (HTB-176), MDA-MB-231 (HTB-26) and PC3 (CRL-1435) were commercially obtained and cultured as recommended by the supplier (ATCC). Cultured cells were re-suspended in culture media containing Hoescht 33342 (Sigma) and incubated at 37°C for 60 minutes prior to sorting.

Cultured cells were flow-sorted to isolate individual cells from each of the cell lines according to phase (G0/G1, M/G2 and S). Cells were isolated and sorted using the FACSJazz™ (Becton Dickinson) at 500 events per second using a 100 micron nozzle. Single cells were defined by gating on forward and side scatter area/width. Phase was inferred from Hoescht 3342 DNA-fluorescent dye, then cells were individually deposited and lysed in wells of a 96-well PCR plate containing 3uL of Cells-to-Ct™ lysis buffer (Life Technologies). The proportion of cells in G0/G1 phases varied from 54% of PC-3 cells to 73% of H9 cells (Supplementary Figure S1).

Genes Assayed

A set of 333 probes was designed. It contained cell cycle associated genes and provided coverage of the entire cell cycle based on peak expression and periodicity information derived from an integrated database of cell cycle expression profiling experiments [16]. Non-cell cycle associated genes had primary roles in the inflammatory response and included housekeeping controls without a Cyclebase ranking. Genes with a Cyclebase ranking < 1000 were placed in the *ranked* set (n = 119)

and all other probes were considered part of the *unranked* set (n=214).

cDNA Conversion and Multiplexed Target Enrichment (MTE)

After lysis, RNA was converted to cDNA with SuperScript VILO (Life Technologies). Primers for 333 genes were pooled and cDNA was enriched in a multiplexed amplification (MTE) reaction according to the nCounter Single Cell Expression protocol (NanoString). The MTE samples were hybridized overnight at 65°C with an nCounter CodeSet containing probes for all enriched targets (cell cycle related, unrelated genes and controls) and internal controls as recommended by the manufacturer.

Statistical Analysis

Dichotomization and Thresholding

The nCounter Analysis System reports the number of counts of each observed nucleic acid target. We transformed the counts with a shifted log-2 transformation so that $lCount = \log_2(\text{count}+1)$. In examining histograms of the transformed data, $lCount$, we found evidence of bi-modality (e.g. Figures 1C, 2). It has been previously observed [2,9,10] in single cell gene expression that genes may appear "off" in a cell, lacking detectable transcript. Thus we hypothesize that in genes with two clusters apparent, the cluster of smaller $lCount$ might represent background noise without detectable expression, and the cluster of larger $lCount$ might correspond to *bona fide* signal. The distribution of $lCount$ in positive controls, which were added at known concentrations, and negative control probes not occurring in human cDNA, additionally supported this hypothesis (Supplementary Figure S2).

We used an empirical Bayes, model-based clustering procedure to discriminate between signal and noise clusters. Via maximum likelihood estimation, we fitted a Gaussian mixture model to an omnibus of expression in all genes to insure that both signal and noise clusters were initially present. The parameter estimates from the omnibus were then used to form an empirical Bayes estimate of a prior distribution for Bayesian Gaussian mixture models fit to each gene separately. The function *thresholdNanoString* available in *SingleCellAssay* implements our thresholding framework, while mixture models are estimated with the *flowClust* R package [28].

Let $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ be the vector of cluster means and variances for a gene. The Bayesian formulation of the mixture model allowed specifying the variance of (μ_1, μ_2) and the expected value of (σ_1^2, σ_2^2) independently. The prior was determined from the parameters of the initial clustering by employing the function *flowClust2Prior* with *kappa=3, Nt=5*. Then for each gene, maximum *a posteriori* parameters are found, subject to the data for that gene and the prior.

Observations with posterior probability >.5 of belonging to the noise cluster were truncated (set to zero) while observations that more likely belonged to the signal cluster were left unchanged. We denoted the truncated, log-transformed value as the Expression Threshold (*et*) in which the value zero denotes no detectable expression, while positive values correspond to increasing values of log-expression. We model the zero value specially and separately from positive values.

Normalization

Normalization was done in a stepwise manner. First data were split by experimental batches and preliminary signal and noise clusters for each gene were estimated using the above thresholding technique. Then the signal and noise peaks were aligned across plates by

subtracting plate-specific intercepts estimated via linear regression, adjusting for cycle effects, akin to non-pooled version of [29]. The normalized log counts were translated so that the minimum normalized value in each gene was zero. Finally, the normalized data were thresholded jointly to produce the data set used for filtering and testing.

Filtering

We adopted a previously published filtering approach [9] based on a robust z-scoring, removing wells with no expression or otherwise outlying in the number of transcripts expressed. We removed non-variable genes (i.e. detectable expression (< 1%) in any cell line). 253 of 333 probes passed these filtering criteria and were carried forward in the analysis. After thresholding and filtering we found that the frequency of expression (the rate at which $et > 0$ in a gene) varied considerably between genes, with a range of .08-.99 and median value of .72.

Amplification Efficiency

We found in examining principle component plots that the first axis of variation corresponded to the number of genes expressed in a well. In cell *k* and gene *g*, let et_{kg} be the thresholded $\log_2$ count, and lCount_{kg} be the un-thresholded $\log_2$ count. Then we defined

$$x_k = 1/N_g \sum_{g=1}^{N_g} 1_{[et_{kg}>0]},$$

where there are N_g genes total, giving the proportion of genes in the panel that were expressed in a cell. We considered several factors before deciding that x_k corresponded to technical variation that should be removed. First, higher-order axes of variation corresponded to identifiable biological factors (e.g., phase, cell line). Orthogonality of x_k to the biological axes of variation suggested that this factor was technical in nature[30]. Secondly, many steps in deriving cDNA from live cells could induce technical, cell-specific variation. These steps include incomplete lysis, variation in reverse transcription to generate cDNA and efficiency differences in the multiplexed, amplicon-specific pre-amplification step. Cell-to-cell variability in any of these could appear downstream as a source of variability[31]. Lastly, we have observed similar phenomena in other cDNA-based single cell gene expression experiments, including multiplexed qPCR and single-cell RNA-seq. As x_k contributed variability to our data and appeared to derive from technical rather than biological sources, we chose to adjust for it as a nuisance source of variability.

In fact, x_k is highly correlated to the log-sum of expression

$$s_k = \log_2\left[1/N_g \sum_{g=1}^{N_g} 2^{\mathrm{lCount}_{kg}}\right]$$

which is equivalent to the log-total read count in RNA sequencing experiments (Supplementary Figure S5). Thus correcting for x_k variability can be seen as a form of normalization, as is typically encountered in RNA-seq.

Hurdle models for zero-inflated expression

In single cell gene expression, we have previously found that accounting for both changes in the frequency of expression and shifts in the PEM produces more sensitive measures of differential expression compared to using either the frequency or the positive values alone, or compared to t-tests on the zero-inflated values [9,32]. We sought to extend this framework to any model that permits a likelihood ratio test on parameters, *e.g.*,

generalized linear or generalized linear mixed models, in order to account for additive cell line effects. Let et_{ijk} denote the expression threshold in the *i*th cell line, *j*th cycle and *k*th cell. Then we model

$$\text{logit}\left(\Pr\left(et_{ijk} > 0\right)\right) = \alpha_i + \beta_j + \gamma_{ij} + x_{\text{ijk}}\delta$$
$$et_{ijk} \mid et_{ijk} > 0 = \alpha_i' + \beta_j' + \gamma_{ij}' + x_{\text{ijk}}\delta' + \epsilon_{ijk}, \quad (1)$$

where α, α' are cell line effects, β, β' are cell cycle effects and γ, γ' are interaction effects between cell line and cell cycle, and ϵ_{ijk} is an independent, normally distributed error. The term $x_{\text{ijk}}\delta$ accounts for cell-to-cell technical variability resulting from variation in reverse transcription and PCR amplification efficiency (see previous section). Jointly modeling the PCR efficiency along with the biological effects of interest is important as one factor can affect the other. Our modeling framework can be extended to regression type models when the right hand side is replaced with a general term $\boldsymbol{X\beta}$ for each component, and even to generalized linear mixed models.

In general, let θ be a vector of parameters for the distribution of $1(et > 0)$ and let θ' be a vector of parameters for $et \mid et > 0$. Then when the distribution of et is divided in this fashion, inference about θ' proceeds conditional on $et > 0$. The log likelihood is then additive in the θ and θ' parameters. Classical hypothesis tests with chi-square asymptotic null distribution, such as Wald or likelihood ratio tests on specific components of θ and θ' are null can be conducted separately. Then the test statistics are added together, combining and summarizing the evidence from the two processes, with the degrees of freedom in the null distribution doubled for the purpose of assigning significance. This approach is dubbed the "Hurdle" model and has been used in economics for several decades [33,34].

Application of Hurdle Model to tests of Cell Cycle Expression Regulation

For each gene, we test whether the cell cycle effect, (β, β'), was equal to zero. The log-likelihoods under both models $M_1 : (\beta, \beta') \neq 0$ and $M_0 : (\beta, \beta') = 0$ are compared. Let λ_0 and λ_1 be -2 times the log-likelihood under models M_0 and M_1, respectively. Then $\lambda_{\text{cycle}} = \lambda_0 - \lambda_1$ gives Wilks' likelihood ratio statistic, and in large samples, the null hypothesis of no cycle effect can be tested by comparing λ_{cycle} to a chi-square distribution with four degrees-of-freedom, as there are three cycles, but with a linear constraint, hence two degrees in each of the discrete and continuous et components.

Proportion of Deviance Explained by Cell Cycle

In order to calculate the proportion of deviance explained by cell cycle, we compare our Hurdle model given by (1) to the same model where all cell cycle effects are omitted (i.e. $(\gamma, \gamma') = (\beta, \beta') = 0$). Let λ_a be -2 times the log-likelihood under this alternative model. The cell cycle deviance ratio is calculated as $d_{cycle} = (\lambda_a - \lambda_1) / \lambda_a$, directly analogous to the calculation of the coefficient of determination R^2 in linear least squares.

The deflated cell cycle deviance ratio is calculated as $d_{cycle} / \max(d_{cycle})$, where $\max\left(d_{\text{cycle}}\right)$ = .27 and occurs in gene TOP2A.

Network Estimation

We extend the conditional, neighborhood-based algorithm of Meinshausen-Bulmann [35] to estimate co-expression networks using the Hurdle model. The standard Meinshausen-

Bulmann algorithm uses L1-penalized regressions to estimate partial correlations between vertices (genes) by treating each vertex as a dependent variable in a regression that includes all other vertices as independent variables. If the vertices are jointly Gaussian, non-zero coefficients correspond to statistical dependences between vertices, conditional on all other factors and so reflect a Gauss-Markov Random Field. Here, since the distribution of expression in single cells is not multivariate Gaussian, edges in our network correspond to conditional correlations (after possible application of the logit link).

Following equation 1, we divide expression into discrete and continuous components, so fit regressions of the form

$$\begin{aligned} \text{logit}\left(\Pr\left(et_g > 0\right)\right) &= \mathbf{X}\beta + \mathbf{et}_{-g}\Gamma_{-g} \\ et_g \mid et_g > 0 &= \mathbf{X}\beta' + \mathbf{et}_{-g}\Gamma'_{-g} + \varepsilon \end{aligned} \qquad (2)$$

where et_g is the expression of the *g*th gene, and $\mathbf{et}_{-g}$ is the matrix of expression of all except the *g*th gene, and **X** is a matrix of cellular covariates. We estimate (β,Γ) and (β',Γ') separately, with distinct L1 penalties λ and λ' for Γ and Γ' using the R package glmnet [36]. Unpenalized parameters β and β' adjust for pre-amplification effect x_k; cell line and cell cycle.

Combining Networks

We connect genes g_1 and g_2 if any one of $\Gamma_{g_1,g_2}, \Gamma_{g_2,g_1}, \Gamma'_{g_1,g_2}, \Gamma'_{g_2,g_1}$, is non-zero at their respective penalties thus take the union of the symmetrized sub-networks. To select the penalty parameters, we fix the number of edges, then find λ and λ' (constant across genes) so that an equal number of edges enter from Γ and Γ'. Other ratios of edges are easily attained by choosing λ and λ' appropriately.

Cellular and Marginal Co-Expression

Even when expression is measured in single cells, co-expression estimates may reflect cluster-specific shifts in mean expression rather than cellular co-expression. Let G_1, G_2 be two genes, and let X be a clustering factor that affects expression of at least one of G_1, G_2. Then an elementary probability calculation shows that

$$\begin{aligned} \text{Cov}(G_1, G_2) &= \text{E}(\text{Cov}(G_1, G_2 \mid X)) \\ &+\text{Cov}(\text{E}(G_1 \mid X), \text{E}(G_2 \mid X)) \end{aligned},$$

so that the unadjusted estimate of covariance $\text{Cov}(G_1, G_2)$ will include marginal shifts in the means $\text{E}(G_1 \mid X), E(G_2 \mid X)$ as well as the average covariance $\text{E}(\text{Cov}(G_1, G_2 \mid X))$. If X is measured, then it can be used to adjust the regressions in equation (2) to remove the effect of shifts in the mean and so isolate the effect of $\text{E}(\text{Cov}(G_1, G_2 \mid X))$.

Acknowledgements

The authors would like to thank Seely Kaufmann and Rich Boykin for bioinformatics support associated with CodeSet design. AM, GF and RG were funded by National Institute of Health grant R01 EB008400.

References

1. Chen Y, Zhong JF (2008) Microfluidic devices for high-throughput gene expression profiling of single hESC-derived neural stem cells. Methods Mol Biol 438: 293–303. doi:10.1007/978-1-59745-133-8_22.

2. Levsky JM, Shenoy SM, Pezo RC, Singer RH (2002) Single-cell gene expression profiling. Science (80-) 297: 836–840. doi:10.1126/science.1072241.

3. Tang F, Lao K, Surani MA (2011) Development and applications of single-cell transcriptome analysis. Nat Methods 8: S6–S11.

4. Kalisky T, Quake S (2011) Single-cell genomics. Nat Methods 8: 311–314. doi:doi: 10.1038/nmeth0411-311.

5. Jensen KB, Watt FM (2006) Single-cell expression profiling of human epidermal stem and transit-amplifying cells: Lrig1 is a regulator of stem cell quiescence. Proc Natl Acad Sci U S A 103: 11958–11963. doi:10.1073/pnas.0601886103.

6. Narsinh KH, Sun N, Sanchez-Freire V, Lee AS, Almeida P, et al. (2011) Single cell transcriptional profiling reveals heterogeneity of human induced pluripotent stem cells. J Clin Invest 121: 1217–1221. doi:10.1172/JCI44635.

7. Buganim Y, Faddah DA, Cheng AW, Itskovich E, Markoulaki S, et al. (2012) Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. Cell 150: 1209–1222.

8. Chen W-C, Wu P-H, Phillip JM, Khatau SB, Choi JM, et al. (2013) Functional interplay between the cell cycle and cell phenotypes. Integr Biol (Camb) 5: 523–534. doi:10.1039/c2ib20246h.

9. McDavid A, Finak G, Chattopadyay PK, Dominguez M, Lamoreaux L, et al. (2012) Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. Bioinformatics 29: 461–467. doi:10.1093/bioinformatics/bts714.

10. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, et al. (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature 498: 236–240. doi:10.1038/nature12172.

11. Cragg J (1971) Some statistical models for limited dependent variables with application to the demand for durable goods. Econom J Econom Soc 39: 829–844.

12. Jones A (1989) A double hurdle model of cigarette consumption. J Appl Econom 4: 23–39.

13. Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, et al. (2008) Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotechnol 26: 317–325. doi:10.1038/nbt1385.

14. Yosef N, Shalek AK, Gaublomme JT, Jin H, Lee Y, et al. (2013) Dynamic regulatory network controlling TH17 cell differentiation. Nature 496: 461–468. doi:10.1038/nature11981.

15. Böhmer RM, Ellwart J (1981) Combination of BUdR-quenched Hoechst fluorescence with DNA-specific ethidium bromide fluorescence for cell cycle analysis with a two-parameter flow cytometer. Cell Tissue Kinet 14: 653–658.

16. Gauthier N, Jensen L, Wernersson R, Brunak S, Jensen T (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. doi:10.1093/nar/gkp1044.

17. Tan RZ, Van Oudenaarden A (2010) Transcript counting in single cells reveals dynamics of rDNA transcription. Mol Syst Biol 6: 358.

18. Naka K, Dansako H, Kobayashi N, Ikeda M, Kato N (2006) Hepatitis C virus NS5B delays cell cycle progression by inducing interferon-beta via Toll-like receptor 3 signaling pathway without replicating viral genomes. Virology 361: 161–173.

19. Zopf CJ, Quinn K, Zeidman J, Maheshri N (2013) Cell-cycle dependence of transcription dominates noise in gene expression. PLoS Comput Biol 9: e1003161. doi:10.1371/journal.pcbi.1003161.

20. Liu D, Ding X, Du J, Cai X, Huang Y, et al. (2007) Human NUF2 interacts with centromere-associated protein E and is essential for a stable spindle microtubule-kinetochore attachment. J Biol Chem 282: 21415–21424. doi:10.1074/jbc.M609026200.

21. Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. Science (80-) 297: 1183–1186.

22. Raj A, van Oudenaarden A (2008) Nature, nurture, or chance: stochastic gene expression and its consequences. Cell 135: 216–226. doi:10.1016/j.cell.2008.09.050.

23. McAdams HH, Arkin A (1997) Stochastic mechanisms in gene expression. Proc Natl Acad Sci U S A 94: 814–819.

24. Boggy GJ, Woolf PJ (2010) A mechanistic model of PCR for accurate quantification of quantitative PCR data. PLoS One 5: e12355. doi:10.1371/journal.pone.0012355.

25. Ma L, Liu J, Liu L, Duan G, Wang Q, et al. (2014) Overexpression of the transcription factor MEF2D in hepatocellular cancer sustains malignant character by suppressing G2/M transition genes. Cancer Res.

26. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, et al. (2002) Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors. Mol Biol Cell 13: 1977–2000.

27. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, et al. (1998) A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. Mol Cell 2: 65–73. doi:10.1016/s1097-2765(00)80114-8.

28. Lo K, Brinkman RR, Gottardo R (2008) Automated gating of flow cytometry data via robust model-based clustering. Cytometry A 73: 321–332.

29. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8: 118–127. doi:10.1093/biostatistics/kxj037.

30. Alter O, Brown PO, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci 97: 10101–10106.

31. Livak KJ, Wills QF, Tipping AJ, Datta K, Mittal R, et al. (2013) Methods for qPCR gene expression profiling applied to 1440 lymphoblastoid single cells. Methods 59: 71–79. doi:10.1016/j.ymeth.2012.10.004.

32. Dominguez MH, Chattopadhyay PK, Ma S, Lamoreaux L, McDavid A, et al. (2013) Highly multiplexed quantitation of gene expression on single cells. J Immunol Methods 391: 133–145. doi:10.1016/j.jim.2013.03.002.

33. Aitchison J (1955) On the Distribution of a Positive Random Variable Having a Discrete Probability Mass at the Origin. J Am Stat Assoc 50: 901–908.

34. Duan N, Manning WG, Morris CN, Newhouse JP (1983) Comparison of for Alternative Care Models for the Demand Medical. J Bus Econ Stat 1: 115–126.

35. Meinshausen N, Bühlmann P (2006) High-dimensional graphs and variable selection with the Lasso. Ann Stat 34: 1436–1462. doi:10.1214/009053606000000281.

36. Friedman J, Hastie T, Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 33: 1–22.

Supplemental Material

Figure S1. Separation of asynchronously cycling cells into three cell cycle phase populations (G0/G1, S and G2/M) via fluorescence activated cell sorting (FACS). H9 (A), MB-231 (D) and PC3 (G) cells were sorted based on DNA content as determined via retention of Hoechst 33342 dye. Individual cells were gated on based on forward scatter versus side scatter for H9 (B), MB-231 (E) and PC3 (H) populations . The number and percentage of H9 (C), MB-231 (F) and PC3 (I) cells in a given phase within the asynchronous population as determined by FACS analysis.

Figure S2. Histogram of log Counts of mRNA for various controls and genes. Positive control primers (for which 100% expression is expected), negative control primers (for which no expression is expected) and two genes with difference expression frequencies are shown. The estimated Gaussian mixture densities from the Empirical Bayes model are superimposed.

Figure S3. P values testing for differential expression in the Hurdle model decompose into discrete and continuous portions. Both discrete and continuous components offer information about differential expression, and combining them permits more sensitive inference.

Figure S4: Ratio of ranked discoveries to total discoveries vs Bonferroni-adjusted P values. A discovery in a ranked gene, as it has been previously found to be cell-cycle regulated, is more biologically plausible than a discovery in an unranked gene. The binomial model uses logistic regression on dichotomized expression values, adjusting for cell line and preamplification efficiency. The hurdle model uses equation (1), while "No cell line, hurdle" uses equation (1) but omits the cell line adjustment. The "raw" and "no cell line, raw" models use ordinary least squares on the log2 expression values, adjusting for preamplification efficiency with and without cell line adjustment (respectively). The Hurdle model achieves the highest number of ranked discoveries per total number of discoveries.

Figure S5: The proportion of expressed genes is related to the log-sum of expression in each cell in our panel of $N_g = 253$ genes.