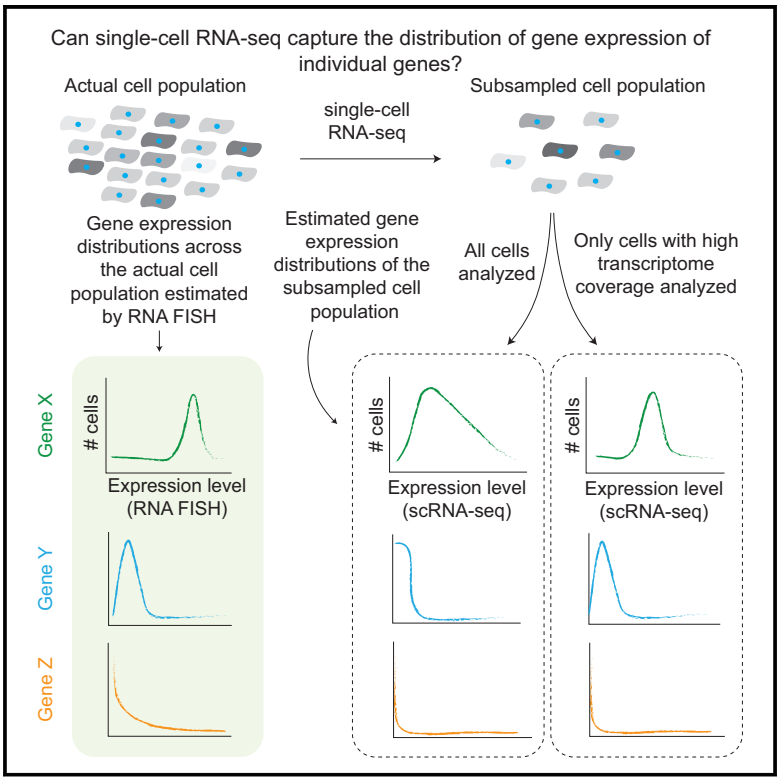
Article

Cell Systems

Rare Cell Detection by Single-Cell RNA Sequencing as Guided by Single-Molecule RNA FISH

Graphical Abstract



Authors

Eduardo Torre, Hannah Dueck, Sydney Shaffer, ..., Junhyong Kim, John Murray, Arjun Raj

Correspondence

jmurr@mail.med.upenn.edu (J.M.), arjunrajlab@gmail.com (A.R.)

In Brief

Single-cell RNA sequencing broadly assays the transcriptome of individual cells, but it is unclear what the trade-offs are when studying the behavior of individual genes. By relying on external controls, we characterize the effect of transcriptome coverage and number of cells analyzed on the accuracy of gene expression distribution estimates.

Highlights

- Estimates of rare cell gene expression vary with transcriptome coverage
- The number of cells analyzed also affects estimates of rare cell gene expression
- In rare cell analysis, cell coverage has a larger effect than the number of cells used
- Internal and external controls guide selection of transcriptome coverage thresholds

Torre et al., 2018, Cell Systems 6, 171–179
February 28, 2018
https://doi.org/10.1016/j.cels.2018.01.014

CellPress

Cell Systems
Article

Rare Cell Detection by Single-Cell RNA Sequencing as Guided by Single-Molecule RNA FISH

Eduardo Torre,[1,7] Hannah Dueck,[3,7] Sydney Shaffer,[1,2] Janko Gospocic,[4,5] Rohit Gupte,[2] Roberto Bonasio,[4,5] Junhyong Kim,[6] John Murray,[3,*] and Arjun Raj[2,3,5,8,*]
[1]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[2]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA
[3]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[4]Department of Cell and Developmental Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[5]Epigenetics Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[6]Department of Biology, University of Pennsylvania, Philadelphia, PA, USA
[7]These authors contributed equally
[8]Lead Contact
*Correspondence: jmurr@mail.med.upenn.edu (J.M.), arjunrajlab@gmail.com (A.R.)
https://doi.org/10.1016/j.cels.2018.01.014

SUMMARY

Although single-cell RNA sequencing can reliably detect large-scale transcriptional programs, it is unclear whether it accurately captures the behavior of individual genes, especially those that express only in rare cells. Here, we use single-molecule RNA fluorescence *in situ* hybridization as a gold standard to assess trade-offs in single-cell RNA-sequencing data for detecting rare cell expression variability. We quantified the gene expression distribution for 26 genes that range from ubiquitous to rarely expressed and found that the correspondence between estimates across platforms improved with both transcriptome coverage and increased number of cells analyzed. Further, by characterizing the trade-off between transcriptome coverage and number of cells analyzed, we show that when the number of genes required to answer a given biological question is small, then greater transcriptome coverage is more important than analyzing large numbers of cells. More generally, our report provides guidelines for selecting quality thresholds for single-cell RNA-sequencing experiments aimed at rare cell analyses.

INTRODUCTION

Single-cell RNA sequencing has emerged as a transformative technology for measuring the transcriptome of individual cells (Kolodziejczyk et al., 2015; Dueck et al., 2016a; Shapiro et al., 2013; Raj and van Oudenaarden, 2008; Symmons and Raj, 2016). The technology has evolved rapidly, and a number of studies comparing technical aspects of the methodologies for single-cell RNA sequencing have emerged recently (Brennecke et al., 2013; Grün et al., 2014; Marinov et al., 2014; Wu et al., 2014; Ziegenhain et al., 2017). These studies compared technical metrics to provide guidelines for the application of single-cell RNA sequencing in general.

The underlying assumption behind these comparisons is that one can conclude which methodology is best purely by comparing technical metrics. Yet, it is clear that whatever the methodology, the inherent challenges of sequencing RNA from a single cell are such that each technique will impose a set of constraints on the data it produces. These constraints, such as the efficiency in capturing the transcriptome of a single cell and the variability in the amount of RNA captured between cells, all affect the resulting accuracy of the putative transcriptome (Figure 1A). Whether these constraints will influence the conclusions drawn from an experiment depends on the specific biological question at hand. Therefore, we believe that the field has reached a point where, instead of relying on metrics devoid of context, we must evaluate the interplay between measurement technique and specific biological context in order to shape our experimental efforts, preferably with external "gold standards" (Grün et al., 2014) to provide robust validation.

Here, we present a case study in single-cell analysis, which evaluates the trade-off between number of cells analyzed, the depth to which each cell is sampled, and our ability to accurately recover distributions of single-cell expression patterns, focusing in particular on the identification and characterization of rare deviating cells in an isogenic population. Our previous work used single-molecule RNA fluorescence *in situ* hybridization (smRNA FISH) (Figure 1B) on many tens of thousands of cells to show that in melanoma cell lines, rare cells (that is, 1 in 50–500) express high levels (that is, dozens to hundreds of mRNA transcripts) of particular genes (e.g., *EGFR*, *NGFR*), and that this expression is associated with resistance to targeted therapy in this subset of cells (Shaffer et al., 2017). Here, the smRNA FISH dataset serves as a gold standard distribution against which we compare single-cell RNA-sequencing data and the distribution of gene expression across a population of cells that has a clear biological interpretation.

We performed single-cell RNA sequencing using DropSeq and Fluidigm methodologies and evaluated our ability to detect rare cell expression at varying thresholds of transcriptome coverage (i.e., number of genes detected per cell). We demonstrate

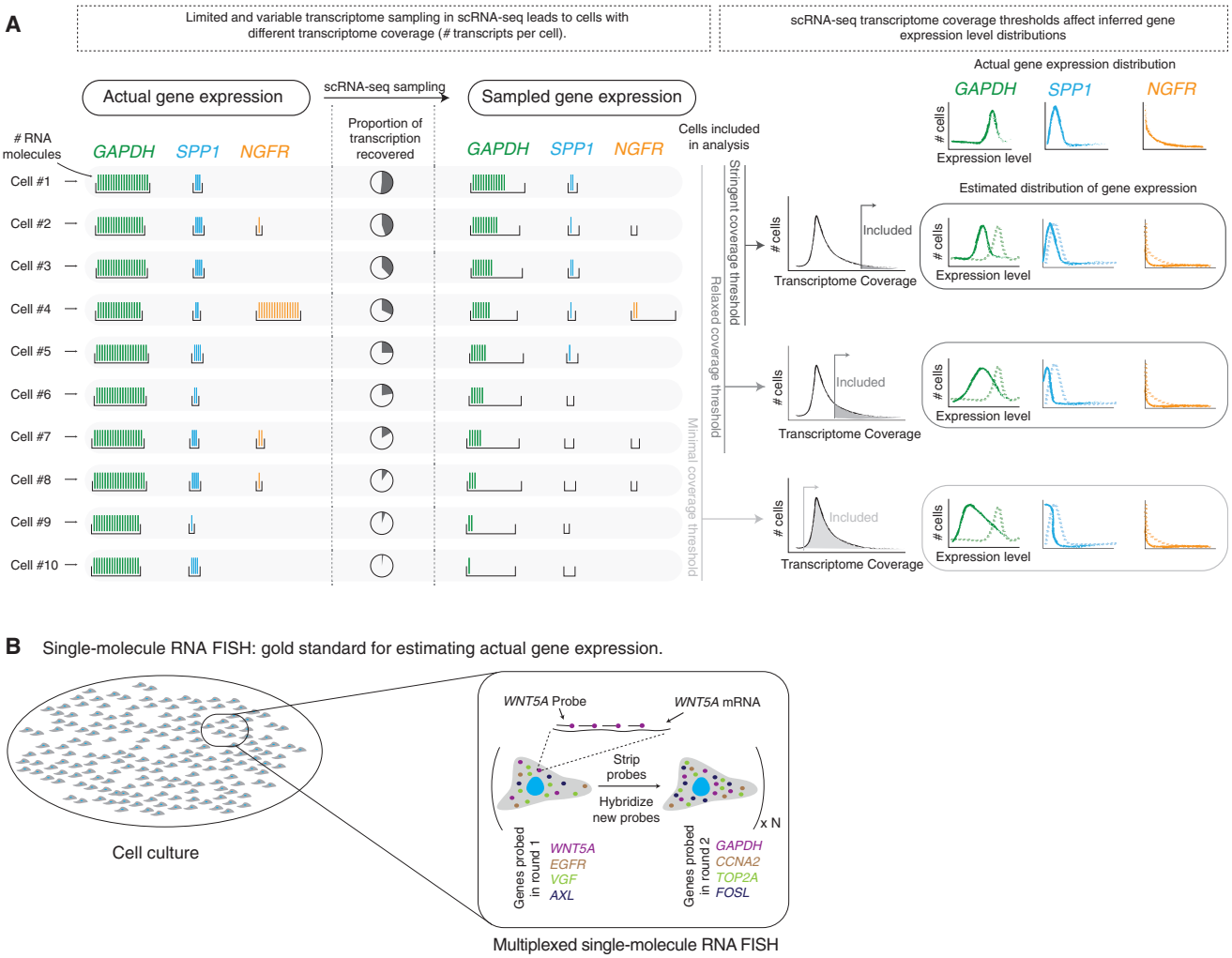


Figure 1. Technical Sampling in Single-Cell RNA Sequencing Can Qualitatively Change Gene Expression Distributions

(A) Single-cell RNA sequencing (scRNA-seq) subsamples the actual transcriptome (left) to an observed transcriptome (middle). Different cells (horizontal rows) can have different degrees of transcriptome coverage. Depending on the number of cells analyzed, the observed expression distribution for any particular gene may not reflect the true distribution (right). We schematically depicted three classes of genes: high, minimally variable expression (*GAPDH*); low, minimally variable expression (*SPP1*); and rare cells with high expression (*NGFR*).

(B) Multiplexed single-molecule RNA FISH is the gold standard for estimating gene expression at the single-cell level. In each round of hybridization, we probe four genes, each with a set of DNA probes containing a common fluorophore. After imaging the resulting RNA spots, we strip the probes and hybridize a new set of probes.

that in many experimental regimes, the apparent distributions of per cell gene expression measured by single-cell RNA sequencing and smRNA FISH are dramatically different, that the observed distribution of cells in different states is dependent on transcriptome coverage, and that below an empirically established threshold of transcriptome coverage, single-cell RNA sequencing cannot reliably separate out genes with true rare cell expression from genes that were just poorly detected. Our approach provides an example of how to apply single-cell sequencing techniques when single-cell analysis is of the essence and technical demands are high due to the rarity of the behavior studied. We suggest that as the field begins evaluating techniques for large-scale data collection, it is a good time to consider the biological context and the requirements it imposes on both data generation and interpretation.

RESULTS

We performed single-cell analysis on a melanoma cell line, WM989-A6, a clonal isolate of the cell line WM989. This cell line serves as a model for melanoma therapy resistance: upon treatment with the BRAF inhibitor vemurafenib, a subset of cells (around 1 in 2,000–5,000) continues to grow in the face of drug. In Shaffer et al. (Shaffer et al., 2017), we used multiplexed smRNA FISH (Figure 1B) to quantitatively measure the expression of resistance markers at the single-cell level. We found that while these cells had overall low average expression of resistance markers such as *EGFR, AXL, WNT5A,* and *NGFR*, occasional rare cells (around 1 in 50–500 cells) would express high levels of these genes. These rare cells were far more likely to be resistant to vemurafenib.

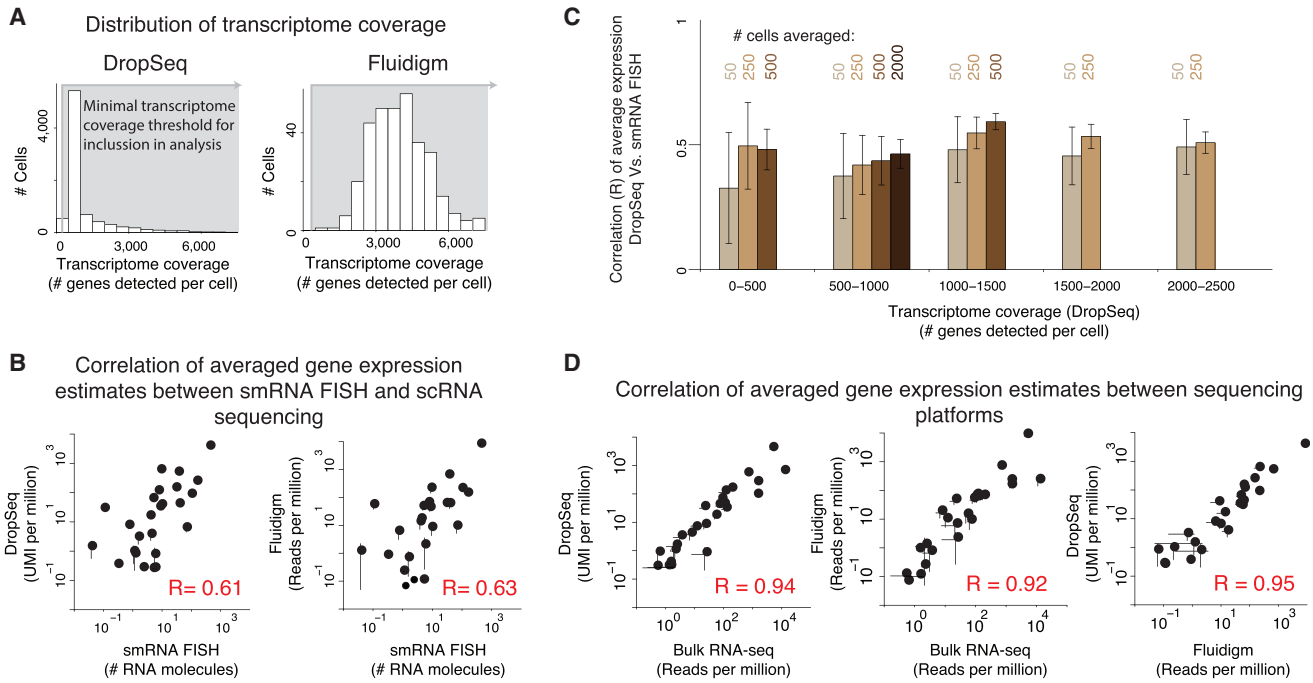


Figure 2. Averaging Gene Expression Estimates across All Cells in Single-Cell RNA Sequencing Shows Good Correspondence across Platforms

(A) Distribution of transcriptome coverage (number of genes detected per cell) for DropSeq (left) and Fluidigm (right).

(B) Correlation of averaged gene expression estimates between single-molecule RNA FISH (smRNA FISH) and single-cell RNA (scRNA) sequencing.

(C) Correlation of average gene expression estimates between DropSeq and smRNA FISH at different levels of transcriptome coverage using four different population sizes (50, 250, 500, and 2000 cells). Error bars in (C) represent ±1 SD across bootstrap replicates.

(D) Correlation of averaged gene expression estimates between sequencing platforms. Error bars in (B) and (D) represent two times the SEM.

To compare our existing smRNA FISH dataset with single-cell RNA sequencing, we used the WM989-A6-G3 cell line with both the DropSeq (Macosko et al., 2015) and the Fluidigm (C1 mRNA Seq HT IFC) platforms. Briefly, the DropSeq platform involves the production of droplets where individual cells lyse and their RNA binds to barcoded RNA-capture beads. The beads are then pooled for library preparation *en masse*. The Fluidigm platform captures cells in microfabricated wells, after which we performed library preparation as per the manufacturer's recommendations, indeed, with the technical support person watching us very closely.

Single-cell RNA sequencing inherently subsamples the transcriptome of each cell because the probability of recovery of any individual transcript is low (typically, only ~10% is recovered; Marinov et al., 2014). Moreover, the degree of subsampling is variable between cells. In principle, subsampling could have very different effects on the interpretation of expression variability from cell to cell depending on the expression level and underlying ground truth distribution. This is schematized on Figure 1A, which focuses on three genes, *GAPDH*, *SPP1*, and *NGFR*, and describes their actual expression and apparent expression after single-cell RNA sequencing. The estimated distributions are all sensitive to the transcriptome coverage threshold, but the effect of thresholding on each distribution is specific to the pattern of expression of the gene. For instance, *GAPDH*, which expresses highly and ubiquitously, is relatively immune to this subsampling: decreasing transcriptome coverage thresholds increases the apparent variability in gene expression across the population, and the population mean might shift, but qualitatively, the distribution is similar. For other low but ubiquitously expressing genes (*SPP1*), however, there can be a qualitative change in which it appears as though they express only in rare cells. Meanwhile, true rarely expressing genes (e.g., *NGFR*) may not be detected at all. This thought experiment illustrates the challenge of imposing a threshold of transcriptome coverage for analysis and of inferring the true distribution of gene expression from single-cell RNA-sequencing data in the absence of a gold standard.

To illustrate these considerations experimentally, we measured the effect of subsampling on transcriptome coverage by observing the number of genes detected per cell (Figure 2A). Effectively, this replaces the pie charts depicted in Figure 1A with experimentally derived values. For our purposes, we defined transcriptome coverage as the number of unique genes detected per cell. To derive these values for DropSeq, we analyzed barcodes from individual beads. As expected, a large number of barcodes had very few reads, potentially due to sources of technical noise such as sequencing errors, barcode synthesis errors, and variability in the number of capture sites per bead (Figure S1A). After we confirmed that most of our data represented transcriptomes from single cells rather than doublets (Figure S1B), we selected the top 8,600 cell barcodes for the remainder of the analysis, with a median sequencing depth of

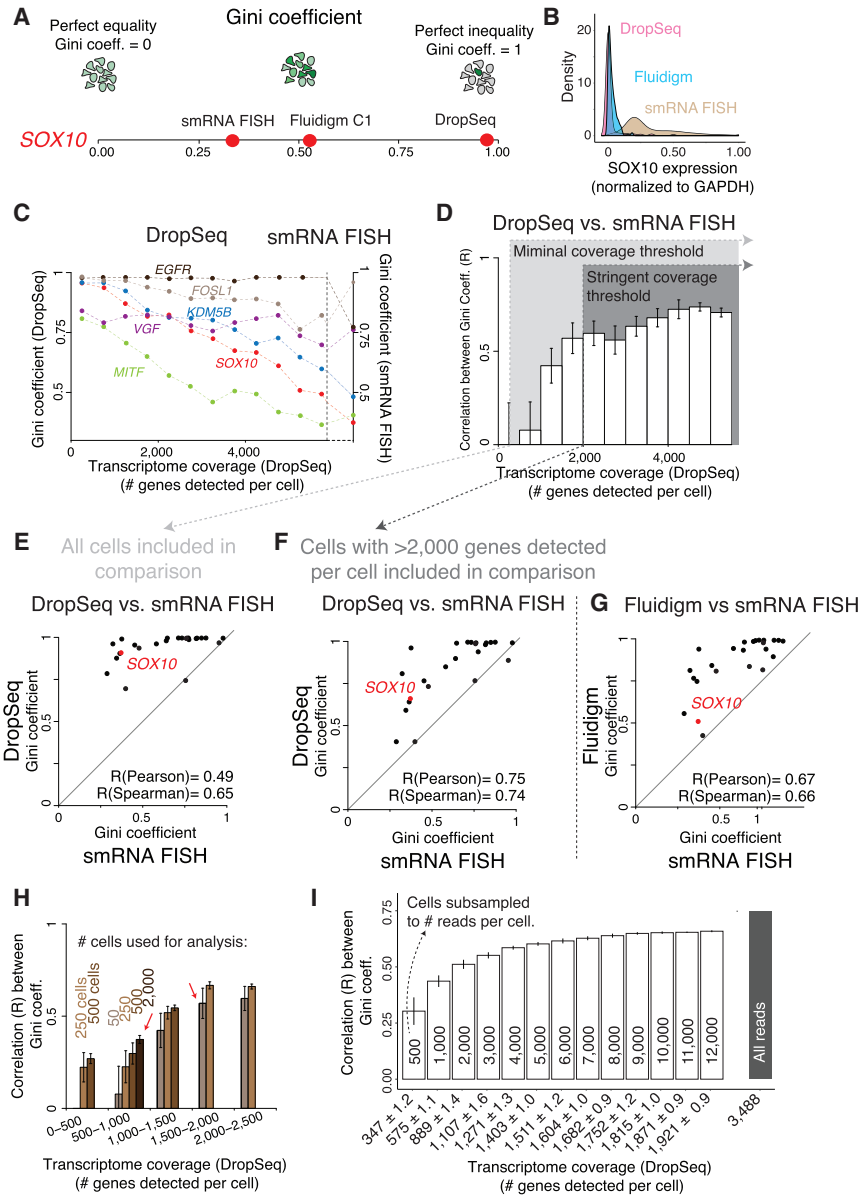


Figure 3. Estimates of Gene Expression Heterogeneity in Single-Cell RNA Sequencing Are Highly Dependent on Transcriptome Coverage

(A) The Gini coefficient measures a gene's expression distribution and captures rare cell population heterogeneity.

(B) Population structure of *SOX10* mRNA levels measured by DropSeq (pink), Fluidigm (blue), and single-molecule RNA FISH (smRNA FISH, brown).

(C) Gini coefficient for six genes measured by DropSeq (left y axis) binned by levels of transcriptome coverage as well as Gini coefficients measured by smRNA FISH (right y axis).

(D) Pearson correlation between Gini coefficients measured through DropSeq and smRNA FISH across different levels of transcriptome coverage (number of genes detected per cell). Error bars represent ±1 SD across bootstrap replicates.

(E and F) Scatterplots of the correspondence between Gini coefficients for 26 genes measured by both DropSeq and smRNA FISH. (E) All cells included in comparison. (F) Cells with >2,000 genes detected per cell included in comparison.

(G) Scatterplot of the correspondence between Gini coefficients for 26 genes measured by Fluidigm and smRNA FISH.

(H) Pearson correlation between Gini coefficient estimates measured by DropSeq and smRNA FISH using different population sizes (number of cells) and levels of transcriptome coverage. Red arrows are discussed in the text. Error bars represent ±1 SD across bootstrap replicates.

(I) Pearson correlation between Gini coefficient estimates measured by DropSeq and smRNA FISH after subsampling cells with high transcriptome coverage to different degrees of read depth. Numbers inside the bars represent the number of reads subsampled. The x axis represents the average number of genes detected across all cells at a given subsample depth. Error bars represent ±1 SD across bootstrap replicates.

6,938 uniquely mapped reads per cell and an interquartile range of 5,553 reads (Figure S1E). Ultimately, we obtained around 8,000 cells with more than 500 genes detected per cell and around 1,100 cells with more than 2,000 genes detected per cell.

Fluidigm produced generally more evenly distributed transcriptomes, albeit with far fewer cells (335 out of a maximum possible of 800). Here, we sequenced the cells to a median depth of ~123,000 uniquely mapped reads per cell with an interquartile range of 110,642 reads (Figures S1C and S1E).

For both platforms, an analysis of sequencing depth suggested that the transcriptome coverages we captured were not limited by the amount of sequencing we performed (Figure S1D). Moreover, in both datasets, we detected ample expression of melanocyte markers and minimal expression of nonmelanocyte markers, providing confidence in the specificity of both datasets (Figure S1F). The wide range of transcriptome coverages in our DropSeq data allowed us to explore the relationship between transcriptome coverage and various metrics of interest.

Out of a variety of metrics, we first looked at one we thought should be relatively insensitive to variability in transcriptome coverage: average gene expression. Despite the varying degrees of transcriptome coverage observed from cell to cell, if the subsampling is unbiased, pooling data from a large number of cells should lead to relative mean expression estimates that are insensitive to specific thresholds and similar across platforms. We pooled single-cell RNA-sequencing data from all cells regardless of transcriptome coverage and compared the resulting mean for each of 23–26 genes with the mean obtained by smRNA FISH (Figure 2B). We found that the correlation with single-molecule RNA FISH was fairly strong for both single-cell RNA-sequencing methods (DropSeq R = 0.61; Fluidigm R = 0.63) (see also Padovan-Merhar et al., 2015;

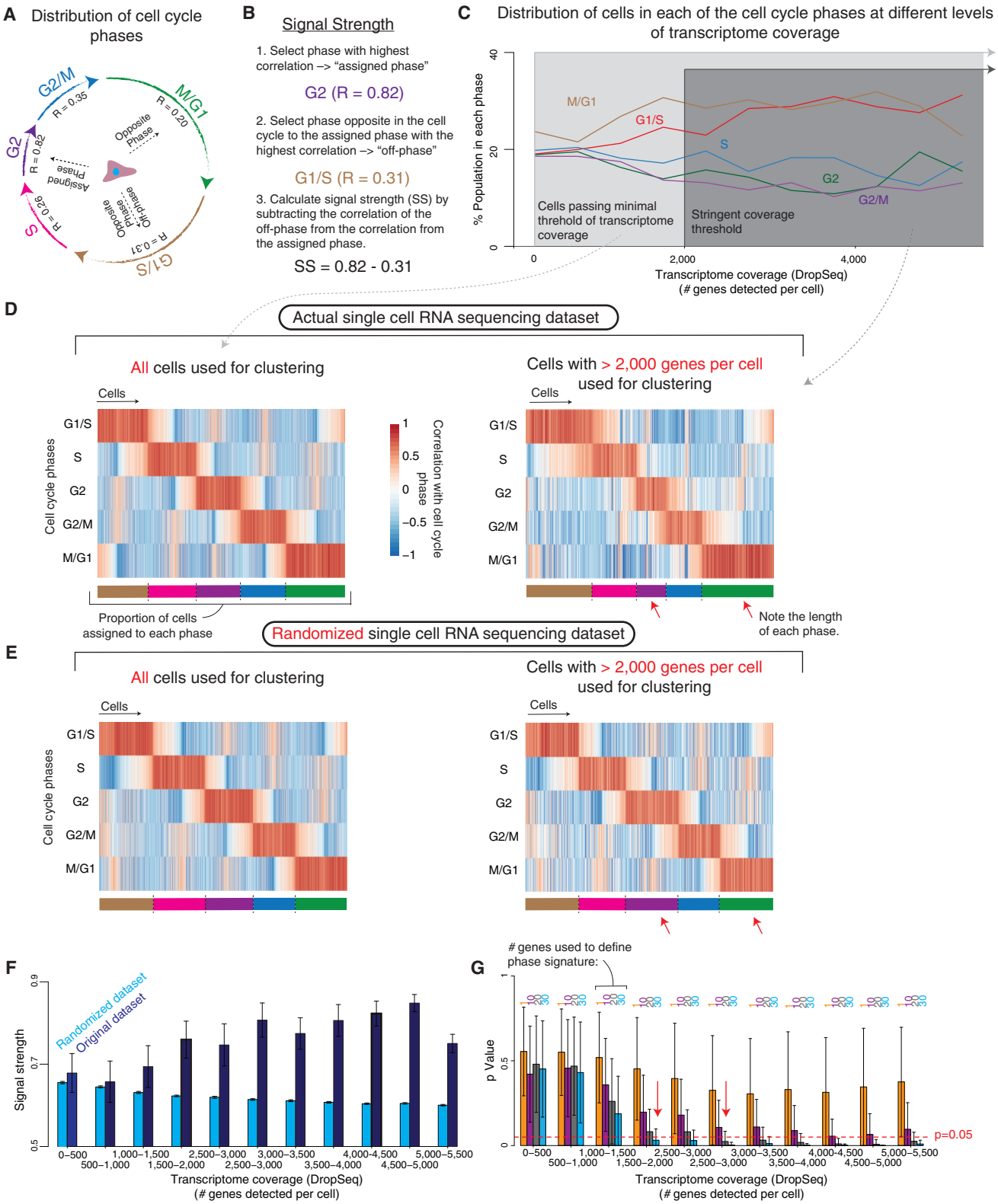


Figure 4. Correct Classification of Single Cells into Multigenic States Is Dependent on Transcriptome Coverage

(A) Schematic depiction of the length of the cell-cycle phases.

(B) Calculation of a cell's signal strength.

(C) Percentage of cells assigned to a cell-cycle phase at different levels of transcriptome coverage (number of genes detected per cell).

(legend continued on next page)

Cabili et al., 2015). Notably, this comparison includes several genes with low average expression due to the rarity of their expression. The reasonable correlation between Fluidigm, DropSeq, and smRNA FISH data demonstrate that in our hands, single-cell RNA sequencing is fairly effective at measuring average expression of even rarely expressed genes. As we expected, as the number of cells included in the analysis increased, so did the correlation between mean gene expression estimates (Figure 2C). However, contrary to our predictions, we found the accuracy of mean gene expression estimates did depend on transcriptome coverage. For example, the estimate obtained from, say, 500 cells with a transcriptome coverage between 1,000–1,500 genes detected per cell yielded a higher correlation than the one obtained from the same number of cells with a shallower transcriptome coverage (e.g., 500–1,000 genes detected per cell). This suggests that subsampling of a cell's transcriptome is nonuniform.

While the correspondence between the single-cell RNA-sequencing data and smRNA FISH was strong enough to capture trends, we wondered whether systematic differences between these approaches might be making the correspondence weaker than it otherwise would be. Therefore, we asked whether the correlation between datasets improves when exclusively comparing sequencing-based techniques. To that end, we compared each single-cell RNA-sequencing dataset with bulk RNA-sequencing data (DropSeq R = 0.94, Fluidigm R = 0.92) and compared Fluidigm and DropSeq with each other (R = 0.95) (Figure 2D). Given the differences in RNA isolation and library preparation, and patterns of coverage between these methods of RNA sequencing, we concluded that the differences between smRNA FISH and single-cell RNA sequencing likely stem from systematic biases in sequencing and not from biases introduced by the different protocols. Accordingly, although the remainder of this study focuses mostly on data generated by DropSeq, we suggest that a similar approach may be taken to characterize Fluidigm data as well.

We next turned to the relationship between transcriptome coverage and the detection of rare cell expression variability. Two aspects of rare cell expression are *a priori* challenging for single-cell RNA sequencing to detect. One is the detection of the rare cell with high levels of expression. The other is the discrimination of genes whose expression is not rare but appears to be rare due to the low capture efficiency of mRNA transcripts (Pierson and Yau, 2015; Dueck et al., 2015; Dueck et al., 2016b). A metric that is able to capture these effects is the Gini coefficient, developed by Corrado Gini as a means of quantifying income inequality. In the context of single-cell expression levels (Jiang et al., 2016), a Gini coefficient of zero signifies an equal distribution of gene expression, whereas a Gini coefficient of one signifies the most extreme level of jackpot expression in which all the RNA is concentrated in a single cell while all the others have none. Intermediate Gini coefficients correspond to intermediate levels of heterogeneity (Figure 3A). (We arrived at similar conclusions using the Kolmogorov-Smirnov statistic; Figures S2A and S2B.) The genes whose expression we analyzed by smRNA FISH had Gini coefficients ranging from 0.29 to 0.98, with housekeeping genes such as *GAPDH* having a Gini coefficient of 0.33, while resistance markers like *EGFR* and *WNT5A* had Gini coefficients of 0.76 and 0.83.

We then wondered how accurate single-cell RNA-sequencing measurements of Gini coefficients would be, given the technical sensitivity of these platforms. We found that when we use very low thresholds for transcriptome coverage the Gini coefficient estimates from single-cell RNA sequencing were generally higher than in smRNA FISH; for instance, *SOX10* has a Gini coefficient of 0.38 by smRNA FISH, but has a Gini coefficient of 0.91 by DropSeq and 0.51 by Fluidigm (Figure 3A). Practically, this means that the population distribution of *SOX10* mRNA levels estimated by single-cell RNA sequencing can be drastically different from the true distribution, as measured by smRNA FISH (Figure 3B).

Given that single-cell RNA sequencing is often plagued by so-called zero inflation, in which some cells artificially have low or zero levels of many transcripts (Dueck et al., 2015, 2016b; Pierson and Yau, 2015), likely inflating a Gini coefficient, we reasoned that accurate estimation of the Gini coefficient may depend on transcriptome coverage. To test this hypothesis, we binned the DropSeq dataset, which had cells of widely varying transcriptome coverages, by the number of genes detected per cell and computed the Gini coefficients for genes for which we had smRNA FISH data (Figure 3C). We found that the Gini coefficient estimates for genes with low variability (e.g., *SOX10*) generally decreased as transcriptome coverage increased, while the Gini estimates for highly variable genes (e.g., *EGFR*) remained high. For each of the bins, we then calculated the Pearson correlation coefficient between the Gini coefficients measured by smRNA FISH and by DropSeq (Figure 3D). We found that keeping only shallow cells yielded virtually no correlation, but cells with progressively deeper transcriptome coverage had increased correlation, with the correspondence increasing sharply until the number of genes detected per cell reached around 2,000.

To see the effect more directly, we imposed two different stringency thresholds for transcriptome coverage, and then asked whether the Gini coefficients for each resultant dataset matched those calculated from smRNA FISH data. The correspondence with the Gini coefficients measured by single-molecule RNA FISH was stronger when keeping only cells in which greater than 2,000 genes were detected (Figures 3E and 3F). Most of the improvement was driven by a drop in

(D and E) Heatmaps representing the correlation of a cell's gene expression signature (columns) with each of the cell-cycle phases (rows) for the DropSeq dataset (D) as well as for a null model (E) where the expression levels of all cycling genes were randomly shuffled within each cell. We analyzed either all cells (left) or only cells with >2,000 genes detected per cell (right). Below each heatmap is a representation of the proportion of cells assigned to each phase of the cell cycle. Notice the length of each bar.

(F) Signal strength across different levels of transcriptome coverage for DropSeq and a null model of randomized DropSeq data. Error bars represent ±1 SD across bootstrap replicates.

(G) The p value of signal strength at different levels of transcriptome coverage using different numbers of genes to characterize the phase. Bar height indicates mean across bootstrap replicates. Error bars represent ±1 SD across bootstrap replicates.

Gini coefficient for genes measured as more ubiquitously expressed by smRNA FISH (Figures 3E and 3F), suggesting that low capture efficiency of mRNA transcripts can artificially raise Gini coefficients when coverage is not sufficiently deep. However, although coverage depth does affect the accuracy of the calculated Gini coefficients, single-cell RNA sequencing ranked genes according to their level of variability similar to smRNA FISH, even at minimal coverage thresholds (Spearman correlations on Figures 3E and 3F). At no point, though, is the correspondence between DropSeq and smRNA FISH perfect, likely reflecting the statistical uncertainty inherent to measuring rare events. This suggests that while single-cell RNA sequencing is able to discriminate qualitatively between variably and uniformly expressed genes, a threshold for transcriptome coverage of 2,000 genes detected per cell (for our DropSeq data) was necessary for reasonably accurate quantification of rare cell expression. The Fluidigm dataset yielded similar correlations (Figure 3G).

This analysis demonstrates that the improvement of Gini coefficient estimates from stringently filtering single-cell RNA sequencing data is driven by decreases in artificially high Gini coefficients: as transcriptome coverage increases, Gini coefficients for genes that are uniformly expressed go down. This leads to the somewhat counterintuitive prediction that having a small number of cells with higher transcriptome coverage leads to more accurate Gini coefficients for rare cell expression than a large number of cells with shallow transcriptome coverage. We tested this prediction by estimating the Gini coefficient for a range of sample sizes (that is, number of cells included per sample) for cells binned by number of genes detected per cell (Figure 3H). We found that increased sample size did improve the similarity of our Gini estimate with the smRNA FISH estimate. However, we also found that using a large number of cells with low transcriptome coverage provided a worse estimate than using a small number cells of higher transcriptome coverage (e.g., compare n = 50 cells with 1,500–2,000 genes detected to n = 2,000 with 500–1,000 genes detected, red arrows in Figure 3H). This is because, while including a large number of cells in the analysis increases the likelihood of detecting rare cells, these large datasets often include many cells with poor transcriptome coverage, which leads to many false-positive rare cell expression events.

Poor transcriptome coverage leads to inaccurate Gini estimates. Given this observation, we wondered if this inaccuracy was simply a product of sequencing depth. In other words, is the difference between cells of high and low transcriptome coverage simply one of number of reads? To simulate the effect of sequencing depth (reads per cell) on the accuracy of Gini estimates, we subsampled cells with high transcriptome coverage (>2,000 genes detected) to various degrees and then calculated the correlation coefficient between the Gini coefficients measured by smRNA FISH and by single-cell RNA sequencing, as well as the number of genes detected at the subsampled depth (Figures 3I and S2C). For both subsampled (Figure 3I) and un-subsampled (Figure 3D) cells, the correspondence of Gini estimates decreased significantly below a coverage of 1,000 genes detected per cell. However, the decrease is more precipitous in our actual dataset, suggesting that sequencing depth does not account for all the differences between cells of high and low transcriptome coverage.

In the analysis described above, smRNA FISH data served as a gold standard with which single-cell RNA-sequencing data could be compared. This comparison defined how coverage depth and number of cells sampled affect the apparent distribution of per cell gene expression; it also allowed us to determine appropriate thresholds of transcriptome coverage when detecting rare cells, our application of choice for this work. Next, we asked whether an appropriate threshold for transcriptome coverage could be identified for a different application in the absence of a smRNA FISH gold standard. Specifically, we asked whether we could use the cell-cycle state to estimate this threshold.

The canonical view of the cell cycle is that at any given point, the number of cells cycling through each of the phases is not uniform: mitosis is short-lived, G1 is not (Figure 4A). So, we expect a population to have many more cells in G1 than in mitosis. We wondered what would be the required transcriptome coverage to recapitulate the expected distribution of cell-cycle phases. To answer this question, we first classified each cell analyzed by DropSeq into a cell-cycle phase (Figures 4C and 4D, left) based on its expression of a panel of genes known to be associated with different phases of the cell cycle (Whitfield et al., 2002). We also created a null expectation of randomly permuted data by shuffling, within each cell, the gene expression values of those genes that mark the different phases of the cell cycle. As expected, after the gene expression values of the cell-cycle marker genes were randomized, the numbers of cells in each of the cycles appear to be relatively equal (Figure 4E, left), in disagreement with the canonical view of the cell cycle. The unshuffled dataset has a similarly equal distribution of cell-cycle phases (Figure 4C, left), suggesting that at minimal thresholds of transcriptome coverage we are unable to accurately classify cells into cell-cycle phases any better than we would a randomly generated dataset.

To assess the relationship between transcriptome coverage and our ability to detect biological signals in the form of cell-cycle phase distribution, we binned cells by the number of genes detected per cell, and then increased the stringency threshold and measured how signal emerged above the randomized control. We defined signal strength as the difference between how well a cell's transcriptome correlated with the signature of the assigned phase and how well it correlated with the "opposite" phases (e.g., for a G2-assigned cell, how well it correlated with G2 minus how well it correlated with G1/S and M/G1 phases) (Figure 4B). Our ability to detect cell-cycle phase improved with the number of genes detected per cell. Moreover, we again found that at a threshold of around 2,000 genes detected, the signal strength significantly increased above our randomized control (Figure 4F), the number of cells in each phase fit more with the canonical view of the cell cycle, with most cells in G1 phase and less in G2 or M (Figure 4C), and the classification of cells differed much more from randomized data, which continued to show a uniform distribution across phases (Figures 4D, right, and 4E, right, see red arrows).

In the example above, we used 31–53 genes to infer cell-cycle position. Next, we defined the relationship between the number of genes used to mark a specific cell-cycle phase and our ability

to classify that phase based on single-cell RNA-sequencing data. We classified cell-cycle phase using random subsets of phase marker genes, for a range of transcriptome coverages (that is, genes detected per cell) (Figure 4G). As expected, our ability to assign cell-cycle phase based on signal strength increased with the number of marker genes available, while more subtle biological signals associated with fewer genes required higher transcriptome coverages to distinguish the signal from randomized data. For example, distinguishing a cell-cycle phase based on the expression of 30 marker genes requires a transcriptome coverage of >1,500 genes/cell, while a cell state defined by 20 marker genes requires >2,500 genes/cell for the detection of any reliable signal (Figure 4G, red arrows).

DISCUSSION

We used a unique complementary set of data to assess how well single-cell RNA sequencing is able to detect high levels of expression in rare cells. Our results suggest that with sufficient transcriptome coverage, single-cell RNA sequencing is able to accurately discern rarely expressing genes from more ubiquitously expressing ones. These results highlight a trade-off inherent to the analysis of single-cell RNA-sequencing data: how deep must the transcriptome coverage of a cell be before including it in the analysis?

Our results suggest that these choices strongly depend on the number of genes important for the biological question under consideration. Thus far, single-cell RNA sequencing has mostly been used for cell-type identification, which involves so many genes that it is relatively robust to low coverage transcriptomes. The rare cell phenotype we examined involved far fewer genes and thus was harder to detect with shallow transcriptomes, as did assignment of cell-cycle phase.

As pointed out by Thomson et al. (Heimberg et al., 2016), the coverage of the transcriptome determines the conclusions one is able to make, and as the number of genes involved in the biology in question decreases, the coverage required will generally increase. For example, cell types typically differ in the expression of thousands of genes, making it relatively easy to discriminate between them even with shallow transcriptomes, even if the cell type is rare (Grün et al., 2015). However, accurately measuring, for example, the cell-to-cell variability in expression of a single gene requires having deep transcriptome coverage to ensure accurate transcript quantification in each cell. In general, most biological processes lie somewhere in between these extremes, and the specifics of the process (e.g., number of genes involved) may impose a particular structure on the data that may or may not be captured at a particular transcriptome coverage.

Given this context, it is important to realize that transcriptome coverage is just one of a number of technical metrics that may be important for evaluating single-cell RNA-sequencing-based approaches to answer any particular biological question. Such considerations will also be important for image-based techniques, where confounders such as cell size (Padovan-Merhar et al., 2015; Battich et al., 2015; Kempe et al., 2015) and others can also affect biological interpretations (Cote et al., 2016). We think that as the field moves toward answering particular biological questions with single-cell RNA sequencing and other single-cell technologies, it will be increasingly important to perform comparative studies, ideally with gold standards, to evaluate the ability to make robust claims.

At the same time, new computational tools such as MAGIC (van Dijk et al., 2017) are under development that aim to recover correlations from shallow-coverage cells in single-cell RNA-sequencing datasets, as well as tools like SAVER (Huang et al., 2017) that are even able to recover distributions of gene expression from shallow-coverage cells. SAVER is able to recover these distributions by training a prediction model across all cells regardless of coverage, thus yielding counts per cell based on a weighted average of the model prediction and the experimental observations. It remains to be seen how much these methods rely on the particulars of the distribution of transcriptome coverage across cells. It also may be that hybrid depth studies, with a smaller subset of cells at very high transcriptome coverage and a large set of cells at shallow transcriptome coverage, may prove useful, with the former discriminating which genes express ubiquitously and the latter finding those that express rarely.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human Melanoma Cells
  - Mouse Cells
- METHOD DETAILS
  - Single Molecule RNA FISH
  - DropSeq
  - Fluidigm C1 mRNA Sequencing
  - Bulk RNA Sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - DropSeq Alignment and Quantification
  - Fluidigm Alignment and Quantification
  - Bulk Sequencing Alignment and Quantification
  - Single Molecule RNA FISH Quantification
  - Selecting Quality Single Cell DropSeq Data
  - Selecting Quality Single Cell Fluidigm Data
  - Sufficiency of Sequencing Depth
  - Tissue-Marker Gene Expression
  - Comparison of Average Measurements
  - Filtering and Normalization for Calculation of Rare Cell Variability
  - Measure of Rare Cell Variability
  - Effect of Library Coverage on Gini Coefficient Estimate
  - Effect of Read Number on Gini Coefficient Estimate
  - Effect of Sample Size (Number of Cells) on Gini Coefficient Estimate
  - Cell-Cycle Phase Classification
  - Effect of Library Coverage on Cell-Cycle Phase Classification
  - Effect of Number of Marker Genes on Cell Cycle Phase Classification
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes two figures and four tables and can be found with this article online at https://doi.org/10.1016/j.cels.2018.01.014.

ACKNOWLEDGMENTS

We thank Ian Mellis for all of his insightful advice, Emily Shields for performing preliminary analysis, and members of the Murray and Raj lab for helpful comments. J.M. and H.D. acknowledge support from the NIH (R21 HD085201). A.R. and E.T. acknowledge support from the NIH New Innovator Award (DP2 OD008514), NIH/NCI PSOC award (U54 CA193417), NSF CAREER (1350601), the NIH (R33 EB019767, P30 CA016520), the NIH (4DN U01 HL129998), the NIH Center for Photogenomics (RM1 HG007743), a Penn Epigenetics Program Pilot award, the Charles E. Kauffman Foundation (KA2016-85223), and the Tara Miller Melanoma Foundation. S.S. acknowledges NIH (F30 AI114475). R.B. acknowledges support from the NIH (DP2MH107055), the Searle Scholars Program (15-SSP-102), the March of Dimes Foundation (1-FY-15-344), a Linda Pechenik Montague Investigator award, and the Charles E. Kauffman Foundation (KA2016-85223).

AUTHOR CONTRIBUTIONS

E.T., H.D., J.M., and A.R. designed the study and wrote the paper. E.T. and H.D. performed the experiments and analysis. S.S. performed RNA FISH experiments. J.G. and R.G. assisted with DropSeq experiments. R.B. and J.K. provided guidance and resources.

DECLARATION OF INTERESTS

A.R. receives consulting income and A.R. and S.S. receive royalties related to Stellaris RNA FISH probes. All other authors declare no competing interests.

Received: July 6, 2017
Revised: October 7, 2017
Accepted: January 17, 2018
Published: February 14, 2018

REFERENCES

Battich, N., Stoeger, T., and Pelkmans, L. (2015). Control of transcript variability in single mammalian cells. Cell *163*, 1596–1610.

Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., et al. (2013). Accounting for technical noise in single-cell RNA-Seq experiments. Nat. Methods *10*, 1093–1095.

Cabili, M.N., Dunagin, M.C., McClanahan, P.D., Biaesch, A., Padovan-Merhar, O., Regev, A., Rinn, J.L., and Raj, A. (2015). Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. Genome Biol. *16*, 20.

Cote, A.J., McLeod, C.M., Farrell, M.J., McClanahan, P.D., Dunagin, M.C., Raj, A., and Mauck, R.L. (2016). Single-cell differences in matrix gene expression do not predict matrix deposition. Nat. Commun. *7*, 10865.

van Dijk, D., Nainys, J., Sharma, R., Kathail, P., Carr, A.J., Moon, K.R., Mazutis, L., Wolf, G., Krishnaswamy, S., and Pe'er, D. (2017). MAGIC: a diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. bioRxiv. https://doi.org/10.1101/111591.

Dueck, H., Khaladkar, M., Kim, T.K., Spaethling, J.M., Francis, C., Suresh, S., Fisher, S.A., Seale, P., Beck, S.G., Bartfai, T., et al. (2015). Deep sequencing reveals cell-type-specific patterns of single-cell transcriptome variation. Genome Biol. *16*, 122.

Dueck, H., Eberwine, J., and Kim, J. (2016a). Variation is function: are single cell differences functionally important?: testing the hypothesis that single cell variation is required for aggregate function. BioEssays *38*, 172–180.

Dueck, H.R., Ai, R., Camarena, A., Ding, B., Dominguez, R., Evgrafov, O.V., Fan, J.B., Fisher, S.A., Herstein, J.S., Kim, T.K., et al. (2016b). Assessing characteristics of RNA amplification methods for single cell RNA sequencing. BMC Genomics *17*, 966.

Grün, D., Kester, L., and van Oudenaarden, A. (2014). Validation of noise models for single-cell transcriptomics. Nat. Methods *11*, 637–640.

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature *525*, 251–255.

Heimberg, G., Bhatnagar, R., El-Samad, H., and Thomson, M. (2016). Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. Cell Syst. *2*, 239–250.

Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J., Raj, A., Li, M., and Zhang, N.R. (2017). Gene expression recovery for single cell RNA sequencing. bioRxiv. https://doi.org/10.1101/138677.

Jiang, L., Chen, H., Pinello, L., and Yuan, G.C. (2016). GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol. *17*, 144.

Kempe, H., Schwabe, A., Crémazy, F., Verschure, P.J., and Bruggeman, F.J. (2015). The volumes and transcript counts of single cells reveal concentration homeostasis and capture biological noise. Mol. Biol. Cell *26*, 797–804.

Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C., and Teichmann, S.A. (2015). The technology and biology of single-cell RNA sequencing. Mol. Cell *58*, 610–620.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell *161*, 1202–1214.

Marinov, G.K., Williams, B.A., McCue, K., Schroth, G.P., Gertz, J., Myers, R.M., and Wold, B.J. (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res. *24*, 496–510.

Padovan-Merhar, O., Nair, G.P., Biaesch, A.G., Mayer, A., Scarfone, S., Foley, S.W., Wu, A.R., Churchman, L.S., Singh, A., and Raj, A. (2015). Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. Mol. Cell *58*, 339–352.

Pierson, E., and Yau, C. (2015). ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. https://doi.org/10.1186/s13059-015-0805-z.

Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A., and Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. Nat. Methods *5*, 877–879.

Raj, A., and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. Cell *135*, 216–226.

Shaffer, S.M., Dunagin, M.C., Torborg, S.R., Torre, E.A., Emert, B., Krepler, C., Beqiri, M., Sproesser, K., Brafford, P.A., Xiao, M., et al. (2017). Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. Nature *546*, 431–435.

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat. Rev. Genet. *14*, 618–630.

Symmons, O., and Raj, A. (2016). What's luck got to do with it: single cells, multiple fates, and biological nondeterminism. Mol. Cell *62*, 788–802.

Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball, C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O., et al. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors. Mol. Biol. Cell *13*, 1977–2000.

Wu, A.R., Neff, N.F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M.E., Mburu, F.M., Mantalas, G.L., Sim, S., Clarke, M.F., et al. (2014). Quantitative assessment of single-cell RNA-sequencing methods. Nat. Methods *11*, 41–46.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative analysis of single-cell RNA sequencing methods. Mol. Cell *65*, 631–643.e4.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
mouse anti-EGFR antibody, clone 225	Millipore	Cat# MABF120; RRID: AB_11212900
anti-mouse IgG-Alexa Fluor488	Jackson Laboratories	Cat# 715-545-150; RRID: AB_2340846
anti-NGFR APC-labelled clone ME20.4	Biolegend	Cat# 345107; RRID: AB_1937276
Critical Commercial Assays		
C1 single cell mRNA seq IFC	Fluidigm	Cat# 101-0063
NEBNext Poly(A) mRNA Magnetic Isolation Module	NEB	Cat# E7490S
NEBNext Ultra RNA Library Prep Kit for Illumina	NEB	Cat# E7530S
Deposited Data		
Raw Single cell RNA seq data	this paper	https://www.dropbox.com/sh/tjdv3mgxle30qiv/AAAXdnYJRZzMeZYaQg7YlUUaa?dl=0 ; GSE99330
Raw Bulk RNA seq data	Shaffer et al., 2017	https://www.dropbox.com/s/ir7fp9ragta8jan/A6_bulk_NoDrug.fastq.gz?dl=0
smRNA FISH data	Shaffer et al., 2017	https://www.dropbox.com/s/om8uq3z3lxnfdtk/fishSubset.txt?dl=0
Experimental Models: Cell Lines		
WM989	Meenhard Herlyn	N/A
NIH 3T3	Raj Lab	N/A
JC4	Raj Lab	N/A
Oligonucleotides		
smRNA FISH Probe sequences	Biosearch Technologies	Table S4
Software and Algorithms		
smRNA FISH image analysis software	Raj Lab	https://bitbucket.org/arjunrajlaboratory/rajlabimagetools/wiki/Dentist
Single cell RNA sequencing analysis code	This paper	https://www.dropbox.com/sh/scmiu1tbrsxupto/AACTL5iWaW-zxRmAbFXnIwnMa?dl=0
Drop-seq_tools-1.0.1	McCaroll Lab	http://mccarrolllab.com/dropseq/
STAR version 2.4.2a		https://github.com/alexdobin/STAR/releases
mRNASeqHT_demultiplex.pl		https://www.fluidigm.com/c1openapp/scripthub/script/2015-08/mrna-seq-ht-1440105180550-2
HTSeq		https://pypi.python.org/pypi/HTSeq

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Arjun Raj (arjunrajlab@gmail.com).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human Melanoma Cells

We obtained WM989 melanoma cells (female) from the lab of Meenhard Herlyn (M.H.) and derived A6 and A6-G3 single cell subclones in our lab. We grew these cells at 37C in Tu2% media (78% MCDB, 20% Leibovitz's L-15 media, 2% FBS, and 1.68mM CaCl2). This cell line was fingerprinted in the lab of M.H. by short tandem repeat profiling using AmpFlSTR Identifiler PCR Amplification Kit (Life Technologies).

Mouse Cells

We grew 3T3 murine cells (male) at 37C in DMEM media (10% FBS, 0.5% pen/strep), and murine JC4 (undetermined sex, single X chromosome) suspension cells at 37C in IMDM media (10% FBS, 2% pen/strep, 50 ng/ml Kit ligand, 2 U/ml erythropoietin, 4.5 x 10^-5 M monothioglycerol). We did not fingerprint these cells. In our analysis they were used only to track rate of doublets in Dropseq.

METHOD DETAILS

Single Molecule RNA FISH

For smRNA FISH we seeded cells in two-well LabTek chambered coverglasses and cultured them to ~50-70% confluency. We performed smRNA FISH and high-throughput microscopy scans as previously described (Raj et al., 2008; Shaffer et al., 2017). In short, we first fixed adherent cells with 4% formaldehyde in PBS for 10 min at RT and permeabilized with 70% EtOH at 4C overnight. We hybridized FISH probes (DNA oligonucleotides conjugated to fluorescent dyes, Table S4) overnight at 37C, washed away unbound probes, and stained DNA with DAPI prior to acquiring a tiled grid of images. Note that in our imaging system, we measured expression in a single z-plane of the cell; thus, the exact numbers for each cell are not total mRNA counts per cell, but rather an amount proportional to the total. For iterative FISH, we stripped DNA probes and hybridized new ones as in Shaffer et. al. (Shaffer et al., 2017). In short, after an initial round of imaging, we removed bound DNA probes using 60% formamide in 2X SSC during a 15 minutes incubation at 37C. We then removed the formamide with three 15-minute PBS washes at 37C. Finally, we washed one last time with wash buffer prior to adding a new set of DNA probes. For more details on how to make buffers for smRNA FISH and on how to carry out the experiments please visit: https://sites.google.com/site/singlemoleculernafish/

DropSeq

We generated single cell suspensions by trypsinizing adherent cells with 0.05% trypsin-EDTA or by harvesting suspensions cells. We passed all cells through a 40 micron filter and diluted them to 100 cells/ul in 0.01% PBS-BSA. We carried out all subsequent steps as detailed by Macosko et. al, protocol v3.1 (http://mccarrolllab.com/dropseq/). In short, we loaded cells in 0.01% PBS-BSA and barcoded beads (chemgenes Barcoded Bead SeqB, cat. No. MACOSKO-2011-10) in lysis buffer onto a droplet generating microfluidic device. After breaking the droplets, we pooled the beads into aliquots of ~60,000, reverse transcribed the RNA captured by the barcoded beads, and digested unbound poly-dT tails via exonuclease treatment. We PCR-amplified STAMPs (2000 beads per reaction), purified cDNA using AMPure beads and quantified the library via Agilent's High Sensitivity DNA Chip. We then tagmented the resulting cDNA with Nextera XT adapters and purified the final library with Ampure beads. We sequenced all libraries using Nextseq 500 with the custom Dropseq read 1 primer described by Macosko et. al.

Fluidigm C1 mRNA Sequencing

To prepare single cell suspensions, we dissociated WM989-A6-G3 cells as above. We immunostained the cells as per Shaffer et. al (Shaffer et al., 2017). Briefly, we incubated cells for 1 hour at 4C with 1:200 mouse anti-EGFR antibody, clone 225 (Millipore, MABF120) in 0.1% PBS-BSA. We then washed twice with 0.1% PBS-BSA and then incubated for 30 minutes at 4C with 1:500 donkey anti-mouse IgG-Alexa Fluor488 (Jackson Laboratories, 715-545-150). We washed the cells again (twice) with 0.1% PBS-BSA and incubated for 10 minutes with 1:500 anti-NGFR APC-labelled clone ME20.4 (Biolegend, 345107). After we washed the cells with 0.1% PBS-BSA and pelleted them, we resuspended them in Tu2%, passed them through a 35 micron filter, and diluted them to a final concentration of ~350 cells per ul in Tu2%. We prepared the samples and sequencing library according to the manufacturer's instructions (https://www.fluidigm.com/products/c1-system). In short, we loaded and captured single cells on Fluidigm's C1 integrated fluidic circuit and inspected the capture chambers via microscopy. We then lysed the cells, barcoded the captured mRNA via RT with a barcoded primer, and amplified the resulting cDNA via PCR. Unlike DropSeq, this protocol uses no unique molecular identifiers to label RNA molecules. After we harvested the amplified cDNA, we tagmented the library using Nextera's XT DNA sample preparation kit (following Fluidigm's version of the protocol), purified the final library using Ampure beads and quantified using Agilent's High Sensitivity DNA Chip. We sequenced the library using a Nextseq 500.

Bulk RNA Sequencing

We sequenced mRNA in bulk from WM989-A6 populations as per Shaffer et. al. We isolated mRNA and built sequencing libraries using the NEBNext Poly(A) mRNA Magnetic Isolation Module and NEBNext Ultra RNA Library Prep Kit for Illumina. We sequenced the libraries either on a HiSeq 2000 or a NextSeq 500 to a depth of approximately 20 million reads.

QUANTIFICATION AND STATISTICAL ANALYSIS

DropSeq Alignment and Quantification

Initial DropSeq data processing was performed using Drop-seq_tools-1.0.1 (http://mccarrolllab.com/dropseq/), and following protocol described in seqAlignmentCookbook_v1.1Aug2015.pdf, accessed from the same site. Data were aligned using STAR version 2.4.2a, downloaded from github on Jan 21, 2016. Data were aligned to reference genome builds hg38 (Human) and mm10 (Mouse), and using reference transcriptome annotations Gencode21 (Human) and Refseq mm10 (Mouse), concatenated with ERCC sequences. Reference transcriptome annotations Gencode21 (Human) and Ensembl mm10 release 83 (Mouse), concatenated

with ERCC annotations. Briefly, reads with low-quality base in either cell or molecular barcode were filtered and reads were trimmed for contaminating primer or poly-A sequence. Sequencing errors in barcodes were inferred and corrected, as implemented by Drop-seq_tools-1.0.1. Uniquely mapped reads, with <= 1 insertion or deletion, were used in quantification. To account for differences in molecule recovery, cell measurements were normalized to UMI per million (UPM).

Fluidigm Alignment and Quantification

Fluidigm sequence data were demultiplexed using mRNASeqHT_demultiplex.pl (https://www.fluidigm.com/c1openapp/scripthub/script/2015-08/mrna-seq-ht-1440105180550-2). Demultiplexed data were processed in the same manner as DropSeq data, with a few modifications: 5' ends of reads were not trimmed, and reads (rather than UMI) were used for quantification. To account for differences in molecule recovery and sequencing depth, cell measurements were normalized to reads per million (RPM).

Bulk Sequencing Alignment and Quantification

We aligned reads to hg19 and quantified reads per gene using STAR and HTSeq.

Single Molecule RNA FISH Quantification

All image analysis was performed as per Shaffer et. al. (Shaffer et al., 2017). We developed a MATLAB analysis pipeline (freely available here https://bitbucket.org/arjunrajlaboratory/rajlabimagetools/wiki/Dentist) that segments nuclei of individual cells using DAPI images. The pipeline then identifies regional maxima as potential smRNA FISH spots and assigns them to the nearest nuclei. We then select a signal intensity threshold for each smRNA FISH channel to differentiate background from smRNA FISH signal and manually curate the dataset to eliminate imaging artifacts that the software recognizes as smRNA FISH signal. We then extract the position of every cell in the scan and the number of RNA molecules for each fluorescent channel. To match cells across subsequent hybridizations, we developed software that shifts cells in the first hybridization to all potential candidates in the subsequent hybridization (Shaffer et al., 2017). It then chooses the best match as the one that minimizes the total distance for nearby cells. We then matched cells by proximity and discarded those cells that did not match uniquely to a nearby cell.

Selecting Quality Single Cell DropSeq Data

Cell barcodes were classified as quality human cells, based on the following criteria: 1) Greater than 80% of species-specific transcripts were assigned to human, and at least 100 species-specific transcripts were available for assignment. 2) The cell barcode was not assigned a synthesis error. The remaining barcodes were filtered to retain the expected number of cells. Based on experiment, we expected 8640 single human cells, and we retained the 8640 cell barcodes with largest read depth (Table S1).

Selecting Quality Single Cell Fluidigm Data

The Fluidigm system allowed cells to be imaged before processing, and for images to be associated with sequencing data. In our automated setup, not all wells were imaged, and well numbers were not captured in images (though column identity on the chip was known). In order to use images to identify quality single cells, we first re-ordered visual annotations to best match read depth observed per well (so that images of empty wells had low depth compared to images with individual or multiple cells). Given re-ordered images, wells were classified as quality single cells if: 1) based on associated image, the well was annotated as containing a good, single cell, and 2) if the cell appeared to be distinct from wells annotated as empty by read depth. Both criteria were required (Tables S2 and S3).

Sufficiency of Sequencing Depth

We wanted to ensure that our metric of library coverage (number of genes observed) did not reflect sequencing depth. To test whether the DropSeq and Fluidigm experiments were sequenced to a sufficient depth, we examined the relationship between experimental read depth and the average number of genes observed in single cells. To do this, we randomly and uniformly subsampled reads from the DropSeq (or Fluidigm) read counts table, for a variety of experimental sequencing depths. We generated 10 random samples at each sequencing depth, and report the average number of observed genes, across cells and sample replicates. We generated random samples for an average depth per cell of 100, 500, and 1000 – 500,000 (step size of 1000) raw reads per cell. (For DropSeq data, our experimental depth allowed testing depths up to 120,000 average raw reads per cell.) To identify the number of reads to subsample from the read count table, given these raw read depths per cell, we calculated the fraction of all sequenced reads that were assigned to the read count table. At each selected experimental depth, we used this fraction of reads to subsample the read counts table. For Dropseq data, 11.2% of raw reads were uniquely assigned to genes in quality cells. In Fluidigm data, 29.3% of reads were uniquely assigned to genes in quality cells.

Tissue-Marker Gene Expression

We selected tissue marker genes for melanocytes, pancreas, heart, and spleen from TIGER (http://bioinfo.wilmer.jhu.edu/tiger/). For each tissue type the genes were selected for analysis based on the expression level in their respective tissue and their presence in both single cell RNA sequencing datasets.

Comparison of Average Measurements

We calculated mean ± 2 SEM for each measurement type. Two genes with smRNA FISH measurements were excluded (*VGF* and *NGFR*) due to difficulty in quantifying the smRNA FISH measurements. One additional gene (*AXL*) was not observed in the Fluidigm data, and was excluded from Fluidigm comparison. Pearson and Spearman correlations were calculated over genes observed in both DropSeq and Fluidigm and were calculated on a $\log_{10}$ scale.

Filtering and Normalization for Calculation of Rare Cell Variability

It has been shown that a portion of molecular variability across single cells is due to cell volume (Padovan-Merhar et al., 2015). To focus on rare cell variability, we normalized smRNA FISH cells to *GAPDH*, using *GAPDH* levels as a proxy for cell volume. Cells with <50 *GAPDH* molecules observed were filtered prior to normalization. For visualization, we scaled normalized values by 400 so that normalized counts were on roughly the same scale as single cell molecule counts. In order that sequencing data remain comparable to smRNA FISH data, we filtered DropSeq and Fluidigm cells with no observed *GAPDH*. We then scaled the (sequencing-depth normalized) DropSeq and Fluidigm data so that the median *GAPDH* level across cells was 400, so that sequencing measurements were on a similar scale to smRNA FISH measurements.

Measure of Rare Cell Variability

Gini coefficients were calculated using the R package ''ineq''.

Effect of Library Coverage on Gini Coefficient Estimate

To test the effect of library coverage on estimates of population statistics, we binned cells by library coverage, using the number of observed genes as our metric of coverage. We used bins ranging from 0 to 5500 observed genes, with a step size of 500 genes. Sample size (the number of cells in a bin) is expected to affect the estimate of the Gini coefficient. We controlled for sample size (number of cells) by randomly subsampling cells within a bin, to reach 50 cells per bin, prior to calculating the Gini coefficient. Random sampling was repeated 100 times per bin. So that normalization was consistent across all random subsamples, we normalized all cells to cellular *GAPDH* level. As previously, smRNA FISH cells with <50 *GAPDH* molecules were excluded, as were DropSeq and Fluidigm cells with no *GAPDH* observed. For each coverage bin, we calculated the Pearson correlation of Gini coefficient estimates calculated on DropSeq data with those calculated using smRNA FISH data. For each bin, we report the average correlation across subsample replicates ± 1 standard deviation.

Effect of Read Number on Gini Coefficient Estimate

To evaluate dependence of Gini coefficient estimates on sequencing depth, we randomly subsampled cells containing > 2,000 genes detected per cell to various read depths (500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 11000, and 12000 reads per cell). We repeated random subsampling ten times at each depth. Then, for those genes for which we had smRNA FISH data, we obtained *GAPDH*-normalized gene expression estimates and Gini coefficients. At each subsampling depth and for each replicate we obtained a pearson correlation of the Gini coefficients between smRNA FISH and single cell RNA sequencing. For each read depth we report the average correlation across all subsample replicates ± 1 standard deviation.

Effect of Sample Size (Number of Cells) on Gini Coefficient Estimate

To evaluate the effect of sample size on Gini coefficient estimates, we repeated the analysis described above for a variety of numbers of cells for each coverage bin.

Cell-Cycle Phase Classification

To assess our ability to detect biological expression patterns using DropSeq and Fluidigm measurements, we assigned cell-cycle phase to individual cells, following the approach used in Macosko et al. (Macosko et al., 2015) and using cell-cycle marker genes identified in Whitfield et al. (Whitfield et al., 2002). The Macosko et al. approach involves the following steps: 1) Marker genes were filtered to exclude genes that do not cycle in melanoma cells. For each set of genes assigned to a particular cell-cycle phase, the average expression profile was calculated across the data set. For each individual gene within that set, the correlation with this average profile was calculated. Genes with correlation <0.3, in either DropSeq or Fluidigm data, were excluded. 2) Depth-normalized read counts were zero-adjusted and $\log_2$ normalized. 3) For each cell and phase, a phase score was assigned by calculating the average normalized value across marker genes for that phase. 4) Phase scores were z-normalized, first across cells within each phase, and then across phases within each cell. 5) Sample phase was assigned to each cell. To do this, a binary score profile was created for idealized cells at each phase and phase transition. The correlation of a cell's normalized score profile with this set of idealized profiles was calculated. A cell was assigned a phase based on the maximum observed correlation.

Effect of Library Coverage on Cell-Cycle Phase Classification

To test the effect of library coverage on cell phase classification, we binned cells by the number of observed genes as described above and, as above, we controlled for sample size (number of cells) by randomly subsampling cells within a bin, to reach 50 cells per bin. For each set of sampled cells, we classified cell-cycle phase as described above. To generate a random expectation for cell-cycle phase categorization, we generated 1000 random counts tables, shuffling counts across cell-cycle marker genes for each

sample. For each table, we proceeded with cell-cycle phase classification as described above. We used the same sets of samples as used in test data, so that each tested population of cells was compared to a biologically and technically matched population with randomized expression profiles.

To summarize the strength of biological signal, we calculated for each cell the best correlation with an idealized phase profile (the assigned phase for the cell) and the best correlation with an idealized phase profile for an ''off'' phase, or a cell-cycle phase that does not neighbor the assigned phase. We report the difference between these correlations. To provide a population-level statistic, we calculate the average strength of biological system for each tested population (each randomly sampled set of 50 cells). To assess the significance of this statistic, we calculated the same statistic for each null (randomly shuffled) population, and report the fraction of times that a signal as large or larger is observed. Finally, we summarize these results across the randomly sampled populations (the random sets of 50 cells), reporting the average ± 1 standard deviation across subsample replicates.

Effect of Number of Marker Genes on Cell Cycle Phase Classification

To evaluate the effect of the number of available marker genes, we repeated the analysis described above for a variety of numbers of genes. We randomly selected *n* marker genes for each phase (*n* from 1 to 30), and then ran the analysis described above on that subset of genes. We repeated this 100 times for each *n*. For randomized data, we selected *n* genes for each phase once, because the identity of the gene has been lost in randomizing the data. (This means each of the 1000 randomized replicates is essentially a different random gene set sample as well.) So, each of the 100 replicates of test data for a given *n* are compared to the same null expectation.

NOTE: All statistical details for a given analysis, including definition of center and dispersion measurements, and exact value of n are detailed in the figures.

DATA AND SOFTWARE AVAILABILITY

- The raw datasets from DropSeq and Fluidigm have been deposited in the GEO repository under accession number GSE99330. They can also be found at: https://www.dropbox.com/sh/tjdv3mgxle30qiv/AAAXdnYJRZzMeZYaQg7YlUUaa?dl=0
- The raw reads obtained through bulk RNA sequencing can be found at: https://www.dropbox.com/s/ir7fp9ragta8jan/A6_bulk_NoDrug.fastq.gz?dl=0
- The code used for analysis throughout the paper can be found at: https://www.dropbox.com/sh/scmiu1tbrsxupto/AACTL5iWaW-zxRmAbFXnIwnMa?dl=0
- The smRNA FISH data file as well as the code used for analysis can be found at: https://www.dropbox.com/sh/g9c84n2torx7nuk/AABZei_vVpcfTUNL7buAp8z-a?dl=0