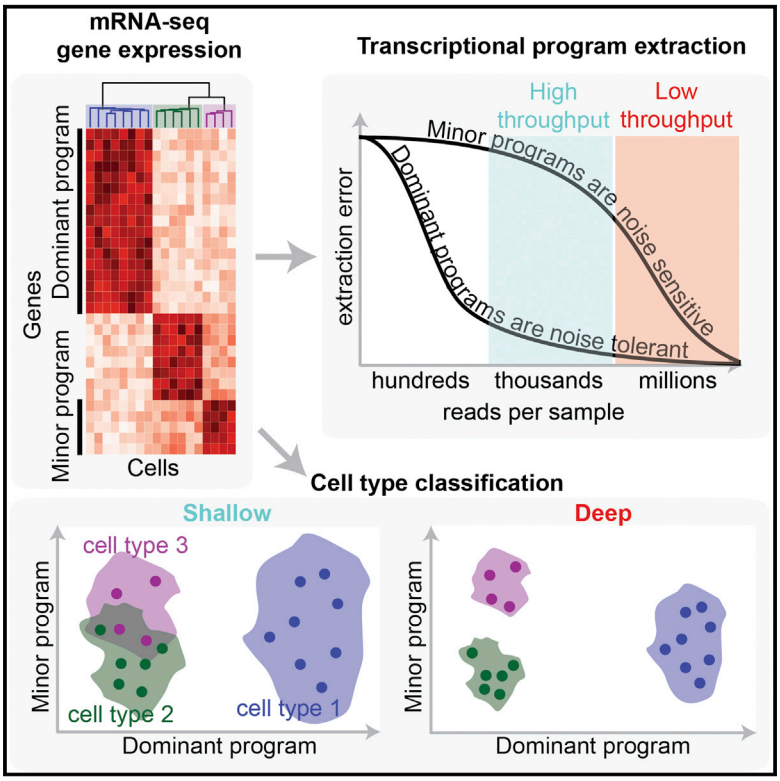
Article

Cell Systems

Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing

Graphical Abstract



Highlights

- Mathematical model reveals impact of mRNA-seq read depth on gene expression analysis
- Modularity in gene expression facilitates robust transcriptional program extraction
- Model suggests dramatic increases in sample multiplexing for many applications
- Read depth calculator determines parameters for optimal experimental design

Authors

Graham Heimberg, Rajat Bhatnagar, Hana El-Samad, Matt Thomson

Correspondence

hana.el-samad@ucsf.edu (H.E.-S.), matthew.thomson@ucsf.edu (M.T.)

In Brief

We develop a mathematical framework that delineates how parameters such as read depth and sample number influence the error in transcriptional program extraction from mRNA-sequencing data. Our analyses reveal that gene expression modularity facilitates low error at surprisingly low read depths, arguing that increased multiplexing of "shallow" sequencing experiments is a viable approach for applications such as single-cell profiling of entire tumors.

Heimberg et al., 2016, Cell Systems 2, 239–250
April 27, 2016
http://dx.doi.org/10.1016/j.cels.2016.04.001

CellPress

Cell Systems
Article

Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing

Graham Heimberg,[1,2,3,4,5] Rajat Bhatnagar,[1,3,5] Hana El-Samad,[1,3,*] and Matt Thomson[3,4,*]
[1]Department of Biochemistry and Biophysics, California Institute for Quantitative Biosciences, University of California, San Francisco, San Francisco, CA 94158, USA
[2]Integrative Program in Quantitative Biology, University of California, San Francisco, San Francisco, CA 94158, USA
[3]Center for Systems and Synthetic Biology, University of California, San Francisco, San Francisco, CA 94158, USA
[4]Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, CA 94158, USA
[5]Co-first author
*Correspondence: hana.el-samad@ucsf.edu (H.E.-S.), matthew.thomson@ucsf.edu (M.T.)
http://dx.doi.org/10.1016/j.cels.2016.04.001

SUMMARY

A tradeoff between precision and throughput constrains all biological measurements, including sequencing-based technologies. Here, we develop a mathematical framework that defines this tradeoff between mRNA-sequencing depth and error in the extraction of biological information. We find that transcriptional programs can be reproducibly identified at 1% of conventional read depths. We demonstrate that this resilience to noise of "shallow" sequencing derives from a natural property, low dimensionality, which is a fundamental feature of gene expression data. Accordingly, our conclusions hold for ~350 single-cell and bulk gene expression datasets across yeast, mouse, and human. In total, our approach provides quantitative guidelines for the choice of sequencing depth necessary to achieve a desired level of analytical resolution. We codify these guidelines in an open-source read depth calculator. This work demonstrates that the structure inherent in biological networks can be productively exploited to increase measurement throughput, an idea that is now common in many branches of science, such as image processing.

INTRODUCTION

All measurements, including biological measurements, contain a tradeoff between precision and throughput. In sequencing-based measurements like mRNA-sequencing (mRNA-seq), precision is determined largely by the sequencing depth applied to individual samples. At high sequencing depth, mRNA-seq can detect subtle changes in gene expression including the expression of rare splice variants or quantitative modulations in transcript abundance. However, such precision comes at a cost, and sequencing transcripts from 10,000 single cells at deep sequencing coverage (10^6 reads per cell) currently requires 2 weeks of sequencing on an Illumina HiSeq 4000.

Not all biological questions require such extreme technical sensitivity. For example, a catalog of human cell types and the transcriptional programs that define them can potentially be generated by querying the general transcriptional state of single cells (Trapnell, 2015). In principle, theoretical and computational methods could elucidate the tradeoff between sequencing depth and granularity of the information that can be accurately extracted from samples. Accordingly, optimizing this tradeoff based on the granularity required by the biological question at hand would yield significant increases in the scale at which mRNA-seq can be applied, facilitating applications such as drug screening and whole-organ or tumor profiling.

The modern engineering discipline of signal processing has demonstrated that structural properties of natural signals can often be exploited to enable new classes of low cost measurements. The central insight is that many natural signals are effectively "low dimensional." Geometrically, this means that these signals lie on a noisy, low-dimensional manifold embedded in the observed, high-dimensional measurement space. Equivalently, this property indicates that there is a basis representation in which these signals can be accurately captured by a small number of basis vectors relative to the original measurement dimension (Donoho, 2006; Candès et al., 2006; Hinton and Salakhutdinov, 2006). Modern algorithms exploit the fact that the number of measurements required to reconstruct a low-dimensional signal can be far fewer than the apparent number of degrees of freedom. For example, in images of natural scenes, correlations between neighboring pixels induce an effective low dimensionality that allows high-accuracy image reconstruction even in the presence of considerable measurement noise such as point defects in many camera pixels (Duarte et al., 2008).

Like natural images, it has long been appreciated that biological systems contain structural features that can lead to an effective low dimensionality in data. Most notably, genes are commonly co-regulated within transcriptional modules; this produces covariation in the expression of many genes (Eisen et al., 1998; Segal et al., 2003; Bergmann et al., 2003). The widespread presence of such modules indicates that the natural dimensionality of gene expression is determined not by the number of genes in the genome but by the number of regulatory modules. By

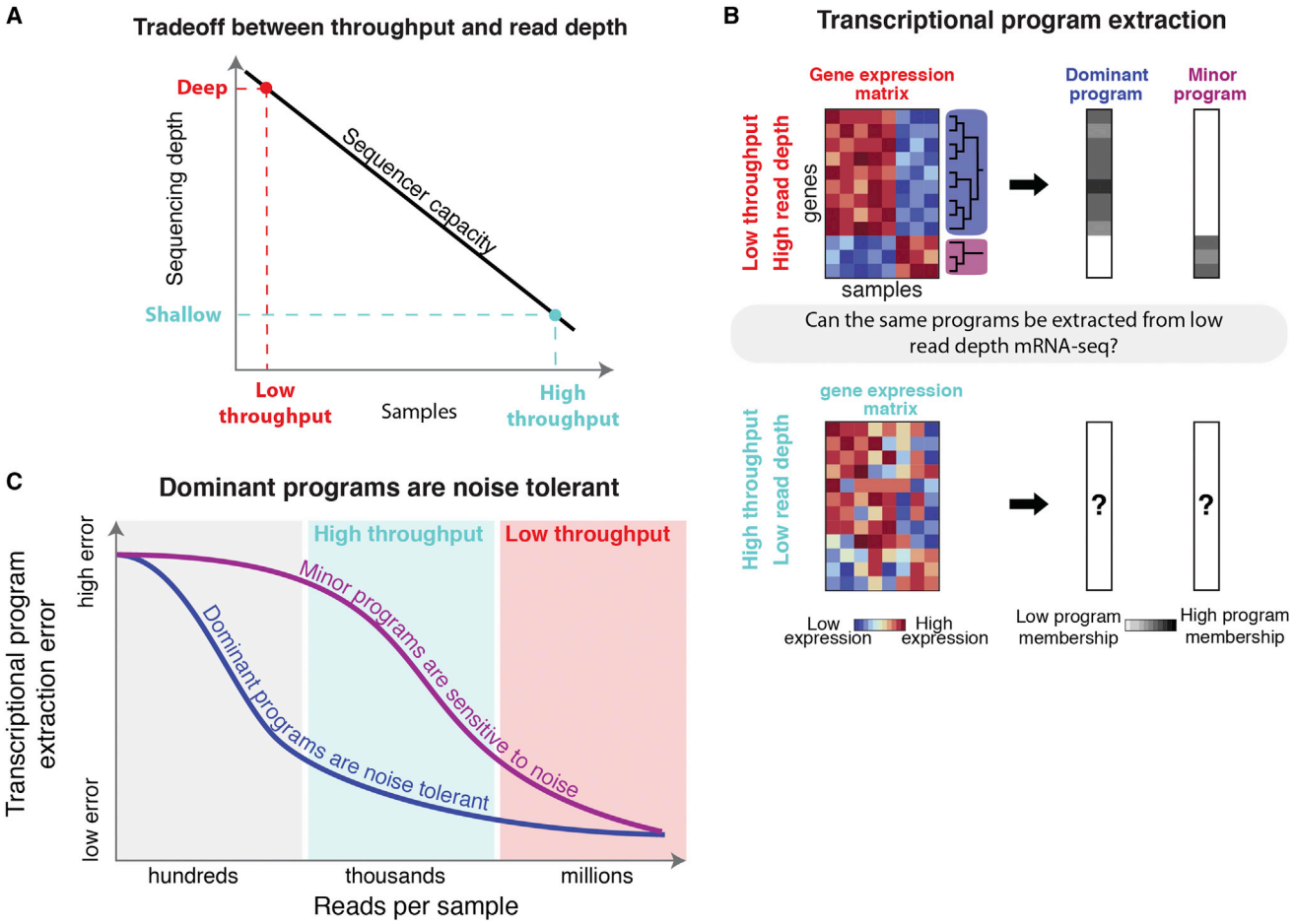


Figure 1. A Mathematical Model Reveals Factors Determining the Performance of Shallow mRNA-Seq

(A) mRNA-seq throughput as a function of sequencing depth per sample for a fixed sequencing capacity.

(B) Unsupervised learning techniques are used to identify transcriptional programs. We ask when and why shallow mRNA-seq can accurately identify transcriptional programs.

(C) Decreasing sequencing depth adds measurement noise to the transcriptional programs identified by unsupervised learning. Our approach reveals that dominant programs, defined as those that explain relatively large variances in the data, are tolerant to measurement noise.

analogy to signal processing, this natural structure suggests that the lower effective dimensionality present in gene expression data can be exploited to make accurate, "inexpensive" measurements that are not degraded by noise. But when, and at what error tradeoff, can low dimensionality be leveraged to enable low-cost, high-information-content biological measurements?

Here, inspired by these developments in signal processing, we establish a mathematical framework that addresses the impact of reducing coverage depth, and hence increasing measurement noise, on the reconstruction of transcriptional regulatory programs from mRNA-seq data. Our framework reveals that "shallow" mRNA-seq, which has been proposed to increase mRNA-seq throughput by reducing sequencing depth in individual samples (Jaitin et al., 2014; Pollen et al., 2014; Kliebenstein, 2012) (Figure 1A), can be applied generally to many bulk and single-cell mRNA-seq experiments. By investigating the fundamental limits of shallow mRNA-seq, we define the conditions under which it has utility and complements deep sequencing.

Our analysis reveals that the dominance of a transcriptional program, quantified by the fraction of the variance it explains in the dataset, determines the read depth required to accurately extract it. We demonstrate that common bioinformatic analyses can be performed at 1% of traditional sequencing depths with little loss in inferred biological information at the level of transcriptional programs. We also introduce a simple read depth calculator that determines optimal experimental parameters to achieve a desired analytical accuracy. Our framework and computational results highlight the effective low dimensionality of gene expression, commonly caused by co-regulation of genes, as both a fundamental feature of biological data and a major underpinning of biological signals' tolerance to measurement noise (Figures 1B and 1C). Understanding the fundamental limits and tradeoffs involved in extracting information from mRNA-seq data will guide researchers in designing large-scale bulk mRNA-seq experiments and analyzing single-cell data where transcript coverage is inherently low.

RESULTS

Statistical Properties of Gene Expression Data Determine the Accuracy of Principal Component Analysis at Low Read Depth

To delineate the impact of sequencing depth on the analysis of mRNA-seq data, we developed a mathematical framework that models the performance of a common bioinformatics technique, transcriptional program identification, at low sequencing depth. We focus on transcriptional program identification as it is central in many analyses including gene set analysis, network reconstruction (Holter et al., 2001; Bonneau, 2008), and cancer classification (Alon et al., 1999; Shai et al., 2003; Patel et al., 2014), as well as the analysis of single-cell mRNA-seq data. Our model defines exactly how reductions in read depth corrupt the extracted transcriptional programs and determines the precise depth required to recover them with a desired accuracy.

Our analysis focuses on the identification of transcriptional programs from mRNA-seq data through principal component analysis (PCA), because of its prevalence in gene expression analysis (Alter et al., 2000; Ringnér, 2008) and its fundamental similarities to other commonly used methods. A recent review called PCA the most widely used method for unsupervised clustering and noted that it has already been successfully applied in many single-cell genomics contexts (Trapnell, 2015). Additionally, research in the computer science community over the last decade has shown that many other unsupervised learning methods, including *k*-means, spectral clustering, and Locally Linear Embedding, are naturally related to PCA or its generalization, Kernel PCA (Ding and He, 2004; Ng et al., 2001; Ham et al., 2004; Bengio et al., 2004). Because of the deep connection between PCA and other unsupervised learning techniques, we expect that our conclusions in this section will extend to other methods of analysis (and we provide such parallel analysis in the Supplemental Information). Here, we focus on PCA because the well-defined theory behind it provides a unique opportunity to understand, analytically, the factors that determine the robustness of program identification to low-coverage sequencing noise.

PCA identifies transcriptional programs by extracting groups of genes that covary across a set of samples. Covarying genes are grouped into a gene expression vector known as a principal component. Principal components are weighted by their relative importance in capturing the gene expression variation that occurs in the underlying data. Decreasing sequencing depth introduces measurement noise into the gene expression data and corrupts the extracted principal components.

If the transcriptional programs obtained from shallow mRNA-seq data and deep mRNA-seq data are similar, then we can accurately perform many gene expression analyses at low depth while collecting data in much higher throughput (Figure 1). We therefore developed a mathematical model that quantifies how the principal components computed at low and high sequencing depths differ. The model reveals that performance of transcriptional program extraction at low read depth is specific to the dataset and even the program itself. It is the dominant transcriptional programs, which capture most variance, that are the most stable.

Formally, the principal components are defined as the eigenvectors of the gene expression covariance matrix, and the principal values λ_i are the associated eigenvalues that equal the variance of the data projected onto the component (Alter et al., 2000; Holter et al., 2001). We use perturbation theory to model how the eigenvectors of the gene expression covariance matrix change when measurement noise is added (Stewart and Sun, 1990; Shankar, 2012). We perform our analysis in units of normalized read counts for conceptual clarity (or normalized transcript counts where appropriate), but an identical analysis and error equation can be derived in FPKM units through a simple rescaling. The principal component error is defined as the deviation between the deep (pc_i) and shallow ($\widehat{pc_i}$) principal components,

$$\|pc_i - \widehat{pc_i}\| \approx \sqrt{\sum_{j \neq i} \left(\frac{pc_i^T \left(\hat{\mathbf{C}} - \mathbf{C} \right) pc_j}{\lambda_i - \lambda_j} \right)^2} \qquad \text{(Equation 1)}$$

where $\mathbf{C}$ and $\hat{\mathbf{C}}$ are the covariance matrices obtained from deep and shallow mRNA-seq data, respectively. Equation 1 can be used to model the impact of shallow sequencing on any given mRNA-seq dataset. Moreover, qualitative analysis of the equation reveals the key factors that determine whether low depth profiling will accurately identify transcriptional programs. As expected, this equation indicates that the principal component error depends on generic features including read depth and sample number, as these affect the difference between the shallow and deep covariance matrices in the numerator of Equation 1 (see the Supplemental Information, section 2.1). However, Equation 1 also reveals that the principal component error depends on a system-specific property: the relative magnitude of the principal values (captured by $\lambda_i - \lambda_j$). Since the principal values correspond to the variance in the data along a principal component, this term quantifies whether the information in the gene expression data is concentrated among a few transcriptional programs. When genes covary along a small number of principal axes, the dataset has an effective low dimensionality, i.e., the data are concentrated on a low-dimensional sub-space, and transcriptional programs can be extracted even in the presence of sequencing noise.

Mouse Tissues Can Be Distinguished at Low Depth in Bulk mRNA-Seq Samples

To understand the implications of this result in the context of an established mRNA-seq dataset, we applied Equation 1 to a subset of the mouse ENCODE data that uses deep mRNA-seq ($>10^7$ reads per sample) to profile gene expression of 19 different mouse tissues with a biological replicate (Shen et al., 2012) (see the Experimental Procedures). The analysis revealed that the leading, dominant transcriptional programs could be extracted with <1% of the studies' original read depth. Specifically, the first three principal components could be recovered with >80% accuracy (i.e., an error of $1 - 0.8 = 20\%$) with just 55,000 reads per experiment (Figures 2A and S1A). To reach 80% accuracy for all of the first nine principal components, only 145,000 reads were needed (Figure S1B). Increasing read depth further had diminishing returns for principal component accuracy. To increase the accuracy of the

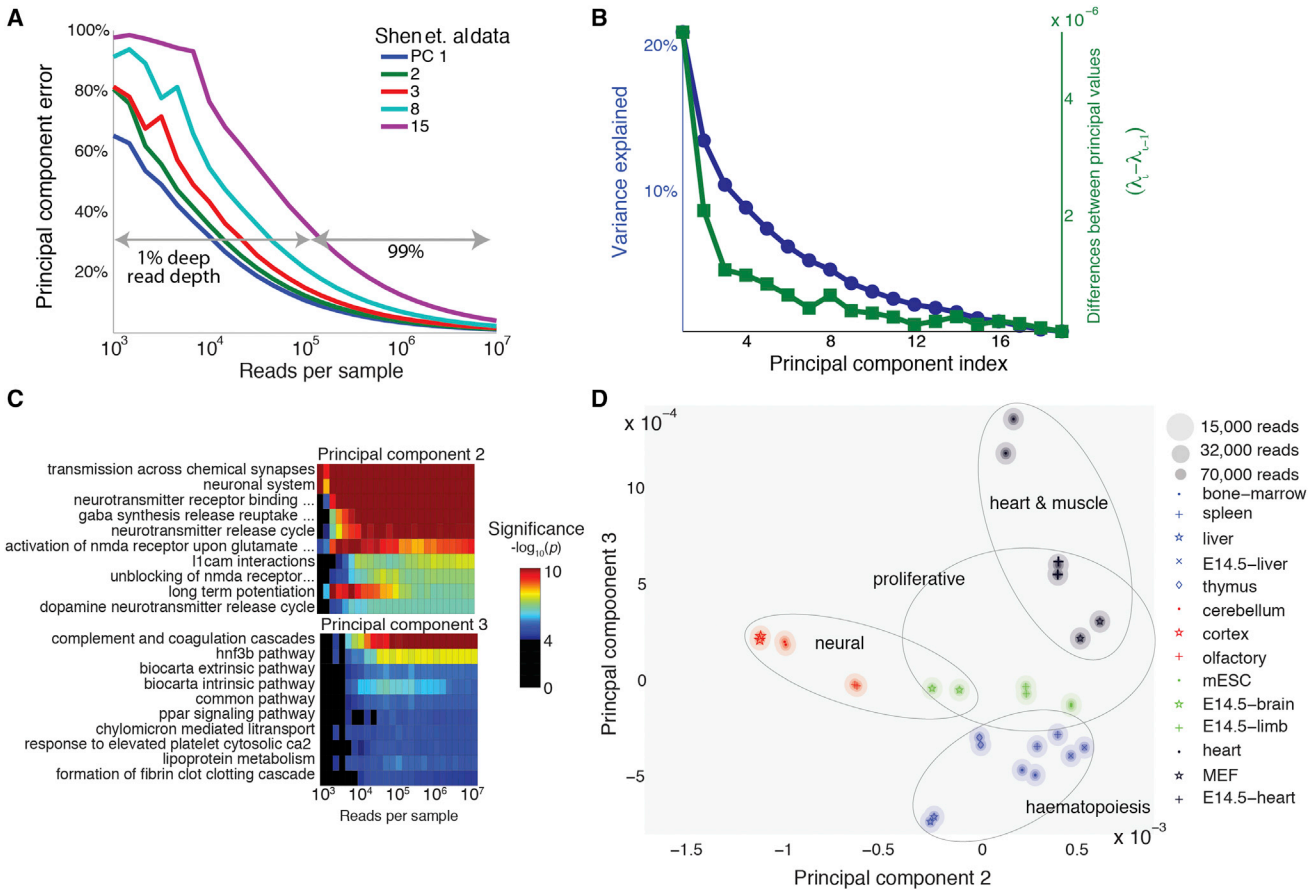


Figure 2. Transcriptional States of Mouse Tissues Are Distinguishable at Low Read Coverage

(A) Principal component error as a function of read depth for selected principal components for the Shen et al. (2012) data. For first three principal components, 1% of the traditional read depth is sufficient for achieving >80% accuracy. Improvements in error exhibit diminishing returns as read depth is increased. Less dominant transcription programs (principal components 8 and 15 shown) are more sensitive to sequencing noise.

(B) Variance explained by transcriptional program (blue) and differences between principal values (green) of the Shen et al. (2012) data. The leading, dominant transcriptional programs have principal values that are well separated from later principal values, suggesting that these should be more robust to measurement noise.

(C) GSEA significance for the top ten terms of principal component two (top) and three (bottom) as a function of read depth. 32,000 reads are sufficient to recover all top ten terms in the first three principal components. (Analysis for first principal component shown in Figure S1C.)

(D) Projection of a subset of the Shen et al. (2012) tissue data onto principal components two and three. The ellipses represent uncertainty at specific read depths. Similar tissues lie close together. Transcriptional program two separates neural tissues from non-neural tissues while transcriptional program three distinguishes tissues involved in hematopoiesis from other tissues. This is consistent with the GSEA of these transcriptional programs in (C).

first three principal components an additional 5% (from 80% to 85%), 55% more reads were required. We confirmed these analytical results by simulating shallow mRNA-seq through direct sub-sampling of reads from the raw dataset (see the Experimental Procedures).

Further, as predicted by Equation 1, the dominant principal components were more robust to shallow sequencing noise than the trailing, minor principal components. This is a direct consequence of the fact that the leading principal values are well separated from other principal values, while the trailing values are spaced closely together. For instance, λ_1 is separated from other principal values by at least $\lambda_1 - \lambda_2 = 5 \times 10^{-6}$, more than two orders of magnitude greater than the minimum separation of λ_{25} from other principal values (1.5×10^{-8}) (Figure 2B). Therefore, the 25th principal component requires almost four million reads, 140 times more than the first principal component, to be recovered with the same 80% accuracy.

To explore whether the shallow principal components also retained the same biological information as the programs computed from deep mRNA-seq data, we compared results from Gene Set Enrichment Analysis applied to shallow and deep mRNA-seq data. At a read depth of 10^7 reads per sample, the first three principal components have many significant functional enrichments with the second and third principal components enriched for neural and hematopoietic processes, respectively (Figure 2C; see Figure S1C for first principal component). These functional enrichments corroborate the separation seen when the gene expression profiles from each tissue are projected onto the second and third principal components (see the Experimental Procedures). Neural tissues (cerebellum,

cortex, olfactory, and embryonic day 14.5 [E14.5] brain) project along the second principal component while the hematopoietic tissues (spleen, liver, thymus, bone marrow, and E14.5 liver) project along the third principal component (Figure 2D).

The statistically significant enrichments of the first three principal components persisted at low sequencing depths. At <32,000 reads per sample, only 0.37% of the total reads, all ten of the top gene sets for these principal components passed our significance threshold of $p < 10^{-4}$ (negative predictive value and positive predictive value in Figures S1D and S1E). To put this result in perspective, using only 32,000 reads per sample (corresponding to PCA accuracies of 81%, 79%, and 75% for the first three principal components, respectively) would allow a faithful recapitulation of functional enrichments while still multiplexing thousands of samples, rather than dozens, in a single Illumina HiSeq sequencing lane. Additionally, this low number of reads was still sufficient to separate the different cell types (Figure 2D). We obtained similar results when working in FPKM units, suggesting that the broad conclusions of our analysis are insensitive to gene expression units (Figures S1F, S1G, and S1H).

Transcriptional States in Single Cells Are Distinguishable with Less Than 1,000 Transcripts per Cell

We wanted to explore whether shallow mRNA-seq could also capture gene expression differences between individual single cells within a heterogeneous tissue, arguably a more challenging problem than distinguishing different bulk tissue samples. In addition to the biological importance of quantifying variability at the single-cell level, single-cell mRNA-seq data provide the necessary context for analyzing the performance of shallow sequencing for two reasons. First, single-cell mRNA-seq experiments are inherently "low-depth" measurements as current methods can capture only a small fraction (~20%) (Shalek et al., 2014) of the ~300,000 transcripts (Velculescu et al., 1999) typically contained in individual cells. Second, since advances in microfluidics (Macosko et al., 2015) now facilitate the automated preparation of tens of thousands of individual cells for single-cell mRNA-seq, sequencing requirements impose a key bottleneck on the further scaling of single-cell throughput.

To probe the impact of sequencing depth reductions on single-cell mRNA-seq data, we analyzed a dataset characterizing 3,005 single cells from the mouse cerebral cortex and hippocampus (Zeisel et al., 2015) that were classified bioinformatically at full sequencing depth (average of ~15,000 unique transcripts per cell) into nine different neural and non-neural cell types. In addition to providing a rich biological context for analysis, this dataset allows for a quantitative analysis of low-depth transcriptional profiling as it incorporates molecular barcodes known as unique molecular identifiers (UMIs) that enable the precise counting of transcripts from each single cell. The Zeisel et al. (2015) data therefore allowed us to analyze the impact of sequencing depth reductions quantitatively in units of transcript counts rather than in the less precise unit of raw sequencing reads.

Similarly to the bulk tissue data, we found that leading principal components in single cells could be reconstructed with a small fraction of the total transcripts collected in the raw dataset. We focused our analysis on three classes of cell types—two classes of pyramidal neurons with similar gene expression profiles and oligodendrocytes—that are transcriptionally distinct. As the first three principal values were well separated from the others (Figure S2A), Equation 1 estimated that the first three principal components could be reconstructed with 11%, 22%, and 38% error, respectively, with just 1,000 transcripts per cell (Figure 3A).

We confirmed this result computationally. With just 100 unique transcripts, we were able to separate oligodendrocytes from the two classes of pyramidal neurons with >90% accuracy. With 1,000 unique transcripts per cell, we were able to distinguish pyramidal neurons of the hippocampus from those of cortex with the same >90% accuracy (Figure 3B). The different depths required to distinguish these subclasses of neural and non-neural cell-types reflect the differing robustness of the corresponding principal components. The first principal component captures a broad distinction between oligodendrocytes and pyramidal cell types (Figure 3C, left) and is the most robust to low read depths. The third principal component captures a more fine-grained distinction between pyramidal neurons but is less robust than the first principal component at low read depth and hence requires more coverage. This is consistent with biological intuition: more depth is required to distinguish between pyramidal neural subtypes than between oligodendrocytes and pyramidal neurons.

We next asked how contributions of individual genes to a principal component change as a function of read depth. For every principal component, we derived a null model consisting of the distribution of the individual gene weightings, called loadings, from a shuffled version of the data (see the Experimental Procedures). Comparing the data to the null model, we found that at a depth of ~340 transcripts, >80% of genes significantly associated with the first principal component could still be detected (Figures 3C and 3D; Experimental Procedures). At just 100 transcripts per cell, we were still able to identify oligodendroycte markers, such as myelin-associated oligodendrocyte basic protein (Mobp) and myelin-associated glycoprotein (Mag), as well as neural markers, such as Neuronal differentiation 6 (Neurod6) and Neurogranin (Nrgn), as statistically significant, and reliably classify these distinct cell types. However, below 100 transcripts per cell, cell-type classification becomes inaccurate, and this is correlated with markers such as Neurod6 being no longer statistically associated with the first principal component.

We were able to reach similar conclusions with three other single-cell mRNA-seq datasets (Shalek et al., 2013; Treutlein et al., 2014; Kumar et al., 2014). With similarly low sequencing depths, we were able to distinguish transcriptional states of single cells collected across stages of the developing mouse lung (Figures S2B–S2D), wild-type mouse embryonic stem cells from stem cells with a single gene knockout (Figures S2E–S2G), and heterogeneity within a population of bone-marrow-derived dendritic cells (Figures S2H–S2J). These results were also not PCA-specific. We additionally examined two of these datasets with t-distributed Stochastic Neighbor Embedding (t-SNE) and Locally Linear Embedding (LLE), two nonlinear alternatives to PCA (Van der Maaten and Hinton, 2008; Roweis and Saul, 2000) and achieved successful classification of transcriptional states (Figures S2K and SKL), in each case recapitulating the results of the original studies with fewer than 5,000 reads per

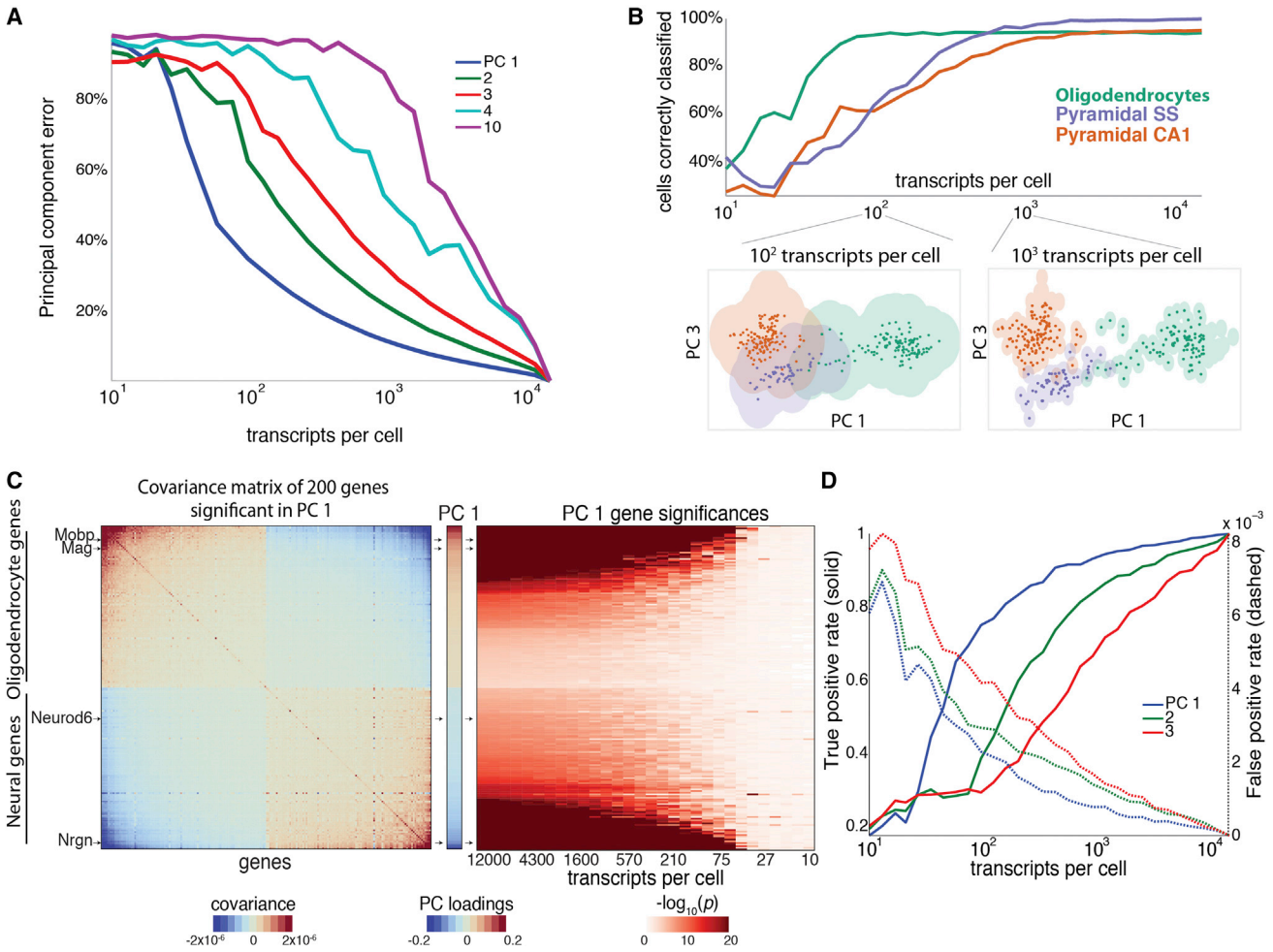


Figure 3. Transcriptional States of Single Cells in the Mouse Brain Are Distinguishable at Low Transcript Coverage

(A) Principal component error as a function of read depth for selected principal components for the Zeisel et al. (2015) data.

(B) Accuracy of cell type classification as a function of transcripts per cell. Accuracy plateaus with increasing transcript coverage. At 1,000 transcripts per cell, all three cell types can be distinguished with low error. At 100 transcripts per cell, pyramidal cells cannot be distinguished from each other, while oligodendrocytes remain distinct.

(C) Covariance matrix of genes with high absolute loadings in the first principal component (left). The genes with the 100 highest positive and 100 lowest negative loadings are displayed. The first principal component is enriched for genes indicative of oligodendrocytes and neurons (middle). Gene significance as a function of transcript count for the first principal component (right).

(D) True and false detection rates as a function of transcript count for genes significantly associated with the first three principal components. Below 100 transcripts per cell, false positives are common.

cell. These results suggest that low dimensionality enables high accuracy classification at low read depth across many methods.

Gene Expression Covariance Induces Tolerance to Shallow Sequencing Noise

In the datasets we considered, the dominant noise-robust principal components corresponded directly to large modules of covarying genes. Such modules are common in gene expression data (Eisen et al., 1998; Alter et al., 2000; Bergmann et al., 2003; Segal et al., 2003). We therefore studied the contribution of modularity to principal component robustness in a simple, mathematical model of gene expression (Supplemental Information, section 2.2). Our analysis showed that the variance explained by a principal component, and hence its noise tolerance, increases with the covariance of genes within the associated module (Figure 4A) and also the number of genes in the module (Figures S3A–S3C). While highly expressed genes also contribute to noise tolerance, in the Shen et al. (2012) dataset we found little correlation between the expression level of a gene and its contribution to the error of the first principal component (R^2 = 0.13; Figure S3D).

This analysis predicts that the large groups of tightly covarying genes observed in the Shen et al. (2012) and Zeisel et al. (2015) datasets will contribute significantly to principal value separation and noise tolerance. To directly quantify the contribution of covariance to principal value separation in these data, we randomly shuffled the sample labels for each gene. In the shuffled data, genes vary independently, which eliminates

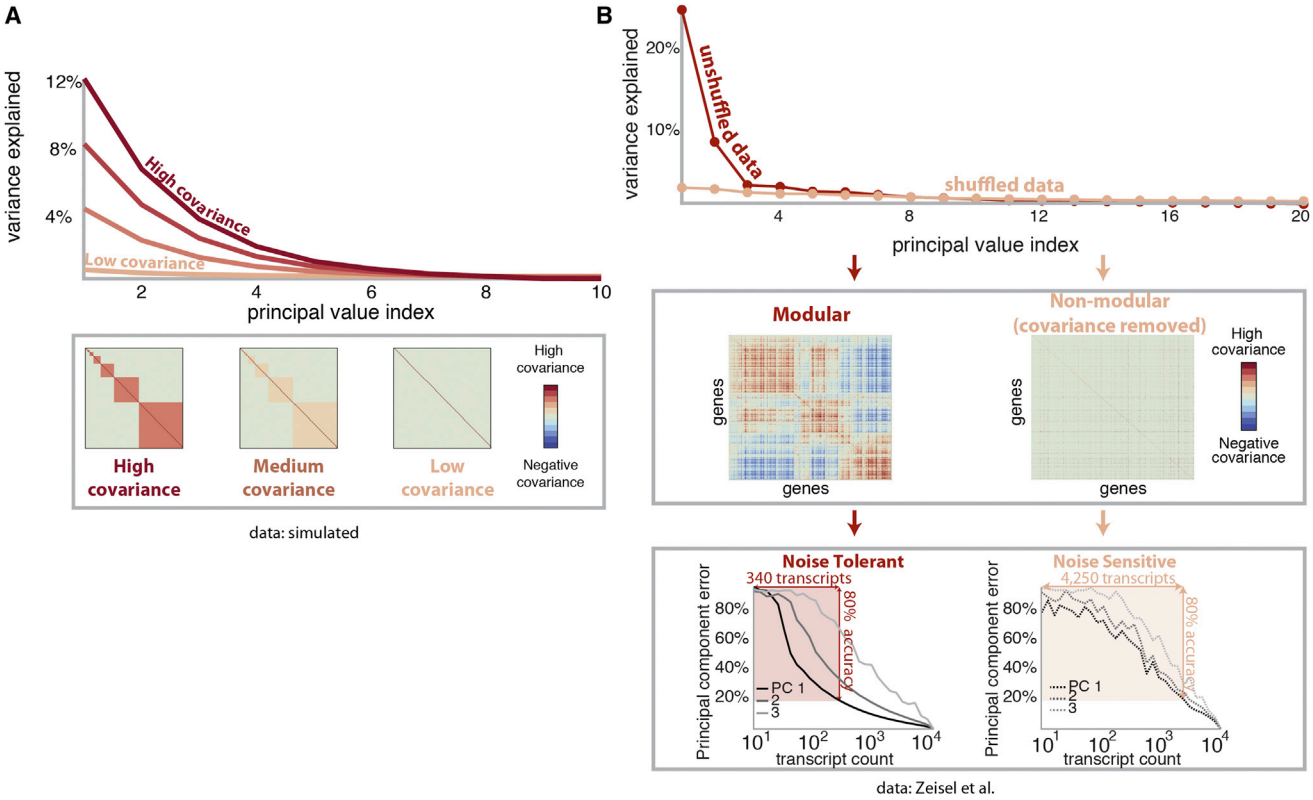


Figure 4. Modularity of Gene Expression Enables Accurate, Low-Depth Transcriptional Program Identification

(A) Variance explained and covariance matrix for increasing gene expression covariance in a model.

(B) Variance explained by different principal components for the Zeisel et al. (2015) dataset. Covariance matrix shows large modules of covarying genes (middle). Dominant transcriptional programs are robust to low-coverage profiling as predicted by model (bottom). Shuffling the dataset destroys the modular structure, resulting in noise-sensitive transcriptional programs. For the shuffled data, 4,250 transcripts are required for 80% accuracy of the first three principal components, whereas 340 transcripts suffices for the original dataset.

gene-gene covariance and raises the effective dimensionality of the data. In contrast to the natural, low-dimensional data, the principal values of the resulting data were nearly uniform in magnitude. This significantly diminished the differences between the leading principal values within the shuffled data (Figure 4B, top).

Consequently, reconstruction of the principal components became more read-depth intensive. For instance to recover the first principal component with 80% accuracy from the shuffled Zeisel et al. (2015) data, 12.5 times more transcripts are required than for the unshuffled data (Figure 4B, bottom). We reached a similar conclusion for the mouse ENCODE data, where shuffling also decreased the differences between the leading principal values and the rest, causing a 23-fold increase in sequencing depth required to recover the first principal component with 90% accuracy (Figure S4).

Large-Scale Survey Reveals that Shallow mRNA-Seq Is Widely Applicable due to Gene-Gene Covariance

Both our analysis of Equation 1 and our computational investigations of mRNA-seq datasets suggest that high gene-gene covariances increase the distance of leading principal values from the rest, thereby enabling the recovery of dominant principal components at low mRNA-seq read depths. This finding, if a common phenomenon, suggests that shallow mRNA-seq may be rigorously employed when answering many biological questions. To assess whether our findings are broadly applicable, we performed a broad computational survey of available gene expression data.

Since both gene covariances and principal values are fundamental properties of the biological systems under study, these quantities may be analyzed using the wealth of microarray datasets available, leveraging a larger collection of gene expression datasets as compared to mRNA-seq (see Figure S5A for analyses of several mRNA-seq datasets). We selected 352 gene expression datasets from the GEO (Edgar et al., 2002) spanning three species (yeast, 20 datasets; mouse, 106 datasets; and human, 226 datasets) that each contained at least 20 samples and were performed on the Affymetrix platform.

Despite the differences between these datasets in terms of species and collection conditions, they all possessed favorable principal value distributions reflecting an effective low dimensionality. For instance, on average the first principal value was roughly twice as large as the second principal value, and together the first five principal values explained a significant majority of the variance, suggesting that these datasets contain a few, dominant principal components (Figure 5A, left). By shuffling these datasets to reorder the sample labels for each gene,

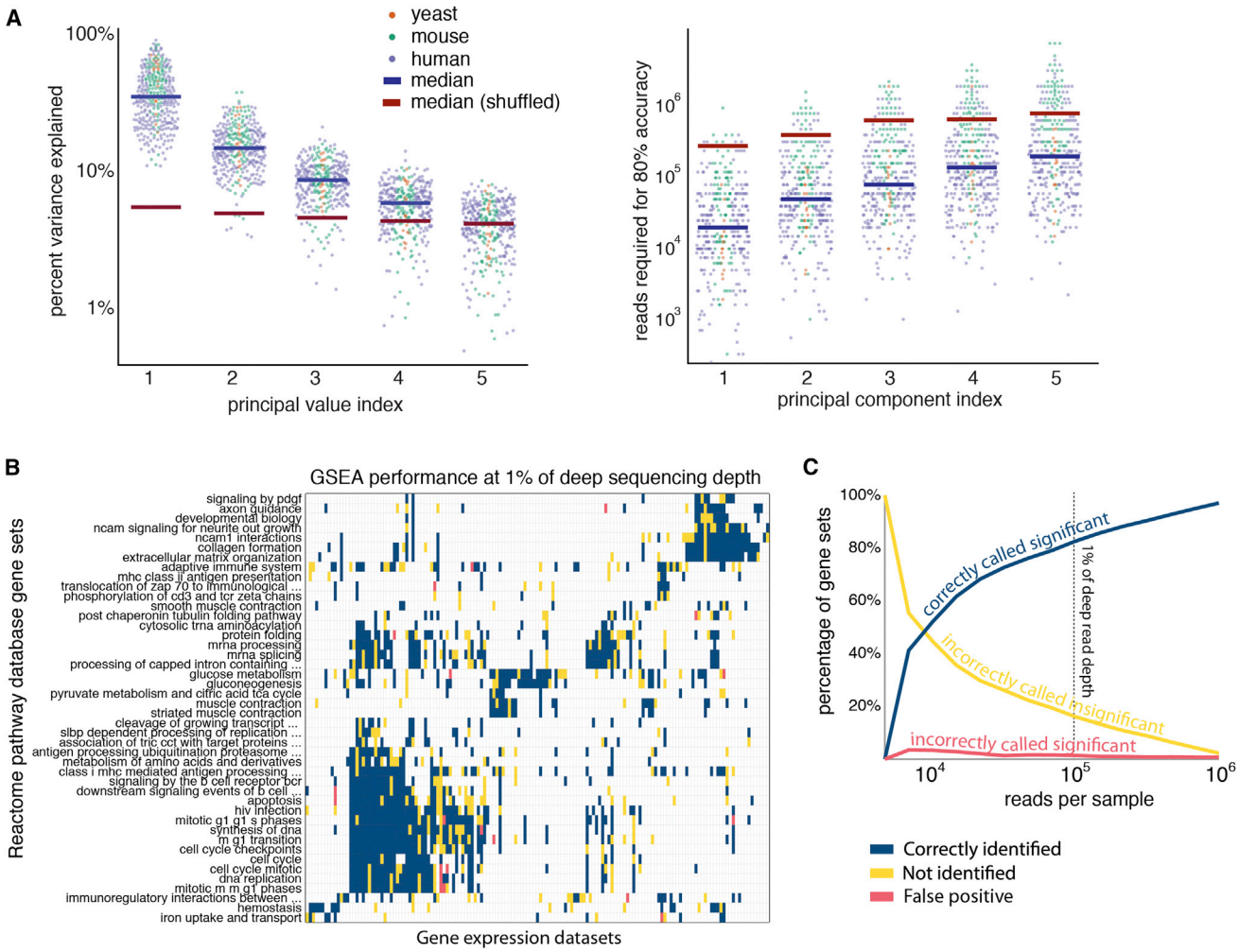


Figure 5. Gene Expression Survey of 352 Public Datasets Reveals Broad Tolerance of Bioinformatics Analysis to Shallow Profiling

(A) Variance explained by the first five transcriptional programs of 352 published yeast, mouse, and human microarray datasets (left). Shuffling microarray datasets removes gene-gene covariance and destroys the relative dominance of the leading transcriptional programs. Read depth required to recover with 80% accuracy the first five principal components of the 352 datasets (right). Removing gene expression covariance from the data requires a median of approximately ten times more reads to achieve the same accuracy.

(B) Accuracy of GSEA of the human microarray datasets at low read depth (100,000 reads, i.e., 1% deep depth). Reactome pathway database gene sets are correctly identified (blue) or not identified (yellow) at low read depth (false positives in red). ~80% of gene sets can be correctly recovered at 100,000 reads.

(C) Accuracy of GSEA as a function of read depth.

we again found that these principal components emerge from gene-gene covariance.

We related this pattern of dominant principal components to the ability to recover biological information with shallow mRNA-seq in these datasets. To generate synthetic mRNA-seq data from these microarray datasets, we applied a probabilistic model to simulate mRNA-seq at a given read depth (see the Experimental Procedures). We found that with only 60,000 reads per sample, 84% of the 352 datasets have $\leq$20% error in their first principal component. This translates into an average of almost 1,000% read depth savings to recover the first principal component with an acceptable PCA error tolerance of 20% (Figure 5A, right). By applying gene set enrichment analysis (GSEA) to the first principal component of each of the 352 datasets at low (100,000 reads per sample) and high read depths (10 million reads per sample), we found that >60% of gene set enrichments were retained with only 1% of the reads (Figures 5B and 5C). This analysis demonstrates that biological information was also retained at low depth.

Collectively, our analyses demonstrate that the success of low-coverage sequencing relies on a few dominant transcriptional programs. We also show that many gene expression datasets contain such noise-resistant programs as determined by PCA and identified them with dominant dimensions in the dataset. Furthermore, low dimensionality and noise robustness are properties of the gene expression datasets themselves and exist independent of the choice of analysis technique. Therefore, unsupervised learning methods other than PCA would reach similar conclusions, an expectation we verified using non-negative matrix factorization (Figure S5B).

The Read Depth Calculator: A Quantitative Framework for Selecting Optimal mRNA-Seq Read Depth and Number of Biological Samples

Because the optimal choice of read depth in an mRNA-seq experiment is of widespread practical relevance, we developed a read depth calculator that can provide quantitative guidelines for shallow mRNA-seq experimental design. Having pinpointed the factors that determine the applicability of shallow mRNA-seq, we applied this understanding to determine the read depth and number of biological samples to profile when designing an experiment. To do so, we simplified the principal component error described by Equation 1 by assuming that the principal values of mRNA-seq data are "well separated," i.e., that the ratio between consecutive principal values λ_{i+1}/λ_i is small (as defined in the Supplemental Information, section 2.1), an assumption justified by our large-scale microarray survey (see Figures S5C and S5D). These assumptions enable us to provide simple guidelines for making important experimental decisions, for example, choosing read depth, N:

$$N \approx \frac{\kappa^2}{n\lambda_i \| pc_i - \widehat{pc_i} \|^2} \qquad \text{(Equation 2)}$$

where n is the number of biological samples and κ is a constant that can be estimated from existing data (see the Supplemental Information, section 2.1 for a derivation of this equation and its limitations). This relationship can be understood intuitively. First, Equation 2 states that the principal component error decreases with read depth, a consequence of the well-known fact that the signal-to-noise ratio of a Poisson random variable is proportional to $\sqrt{N}$. The read depth also depends on λ_i, which comes from the $\lambda_i - \lambda_j$ term of Equation 1. Finally, the influence of the sample number n on read depth follows from the definition of covariance as an average over samples. (Figure S5E shows that n is approximately statistically uncorrelated with principal values across the microarray datasets.)

Equation 2 has implications for optimizing the tradeoff between read depth and sample number in single-cell mRNA-seq experiments. As principal component error depends on the product of read depth and number of samples, error in mRNA-seq analyses can be reduced equivalently in two ways, by either increasing the total number of profiled cells or the transcript coverage. To illustrate this point, we computationally determined the error in the first principal component of the single cell mouse brain data from Zeisel et al. (2015) as a function of cell number. Consistent with Equation 2, our calculations show that increasing the number of profiled cells reduces error in the first principal component (Figure 6A). Furthermore, we show that with the Zeisel et al. (2015) data, multiple different experimental configurations with the same total number of transcripts can yield the same principal component error. For example, 100,000 transcripts divided between either 50 or 400 cells both yield a principal component error of ~20%. This result is of particular relevance in single-cell experiments because transcript depth per cell is currently limited by a ~20% mRNA capture efficiency, and so cannot be easily increased (Shalek et al., 2014). In such cases, limited sequencing resources might be best used to sequence more cells at low depth rather than allocating sequencing resources to oversampling a few thousand unique transcripts.

Experimentalists can use the read depth calculator to predict requirements for read depth or sample number in high-throughput transcriptional profiling given their desired accuracy based on the statistics of principal value separation in our global survey. Figure 6B shows the reads required for desired accuracies and an assumed principal value for a human transcriptional experiment with 100 samples (typical values for the first five principal values for human are indicated in dashed lines). As an illustration, a hypothetical experiment with a typical first principal value of 1.4×10^{-5} (median principal value from the 226 human microarray datasets) and 100 samples where 80% PCA accuracy is tolerable requires less than 5,000 reads per experiment or less than 500,000 reads in total, occupying less than 0.125% of a single sequencing lane in the Illumina HiSeq 4000.

The predictions from this analytically derived read depth calculator are demonstrably accurate. We compared the analytically predicted number of reads required for 80% PCA accuracy in the first five transcriptional programs to the value determined through simulated shallow mRNA-seq for 226 microarray and 4 mRNA-seq human datasets. We determined κ empirically by fitting 50% of the datasets. Cross-validation with the remaining 50% of the datasets showed remarkable agreement between the analytical predictions and computationally determined values. In these calculations, the analytically predicted number of reads required to reach 80% accuracy deviates from the depth required in simulation by less than 10% (Figure 6C). The read depth calculator is available online (http://thomsonlab.github.io/html/formula.html).

Finally, while we use the first principal component for illustration, Equation 2 can be applied to any principal component, including the trailing principal components. Recent work discusses a statistical method to identify those principal components that are likely to be informative, and this work can be used in conjunction with Equation 2 to pinpoint the relevant principal components and the sequencing parameters needed to estimate them satisfactorily (Klein et al., 2015).

DISCUSSION

Single-cell transcriptional profiling is a technology that holds the promise of unlocking the inner workings of cells and uncovering the roots of their individuality (Klein et al., 2015; Macosko et al., 2015). We show that for many applications that rely on the determination of transcriptional programs, biological insights can be recapitulated at a fraction of the widely proposed high read depths. Our results are based on a rigorous mathematical framework that quantifies the tradeoff between read depth and accuracy of transcriptional program identification. Our analytical results pinpoint gene-gene covariance, a ubiquitous biological property, as the key feature that enables uncompromised performance of unsupervised gene expression analysis at low read depth. The same mathematical framework also leads to practical methods to determine the optimal read depth and sample number for the design of mRNA-seq experiments.

Given the principal values that we observe in the human microarray datasets, our analysis suggests that one can profile tens of thousands of samples, as opposed to dozens, while still being

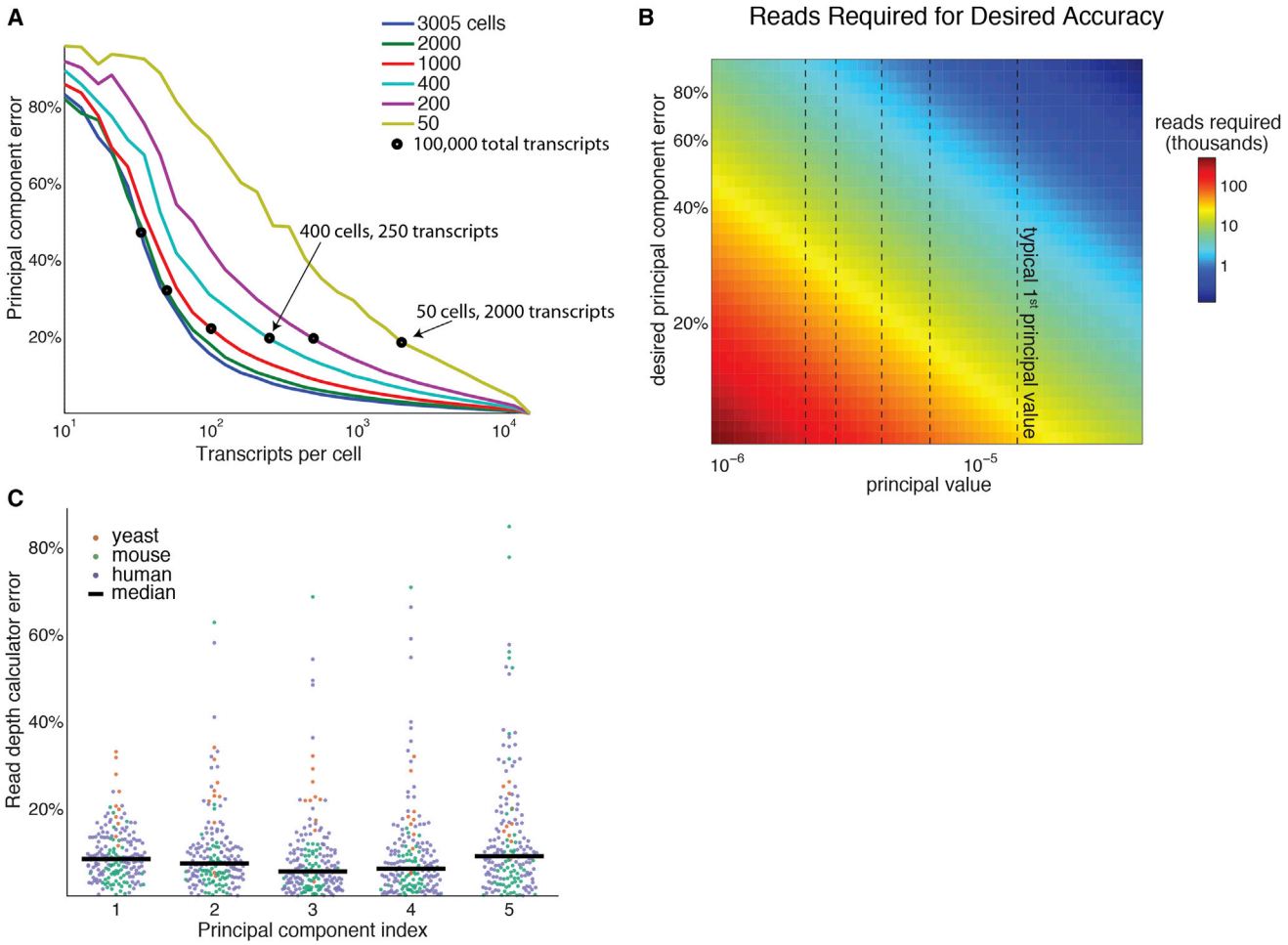


Figure 6. Mathematical Framework Provides a Read Depth Calculator and Guidelines for Shallow mRNA-Seq Experimental Design

(A) Error in the first principal component of the Zeisel et al. (2015) dataset for varying cell number and read-depth. Black circles denote a fixed number of total transcripts (100,000). Error can be reduced by either increasing transcript coverage or the number of cells profiled.

(B) Number of reads required (color) to achieve a desired error (y axis) for a given principal value (x axis). Typical principal values (dashed black vertical lines) are the medians across the 352 gene expression datasets.

(C) Error of the read depth calculator (Equation 2) across 176 gene expression datasets used for validation (out of 352 total). The calculator predicts the number of reads to achieve 80% PCA accuracy in each dataset (colored dots). The predicted values closely agree with simulated results, with the median error <10% for the first five transcriptional programs.

able to accurately identify transcriptional programs. At this scale, researchers can perform entire chemical or genetic knockout screens or profile all ~1,000 cells in an entire *Caenorhabditis elegans*, 40 times over, in a single 400,000,000 read lane on the Illumina HiSeq 4000. Because shallow mRNA-based screens would provide information at the level of transcriptional programs and not individual genes, complementing these experiments by careful profiling of specific genes with targeted mRNA-seq (Fan et al., 2015) or samples of interest with conventional deep sequencing would provide a more complete picture of the relevant biology.

Fundamentally, our results rely on a natural property of gene expression data: its effective "low dimensionality." We observed that gene expression datasets often have principal values that span orders of magnitude independently of the measurement platform and that this property is responsible for the noise tolerance of early principal components. These leading, noise-robust principal components are effectively a small number of "dimensions" that dominate the biological phenomena under investigation. These insights are consistent with previous observations that were made following the advent of microarray technology (Eisen et al., 1998; Segal et al., 2003; Bergmann et al., 2003), proposing that low dimensionality arises from extensive covariation in gene expression. We suggest that the covariances and principal values in gene expression are determined by the architectural properties of the underlying transcriptional networks, such as the co-regulation of genes, and therefore it is the biological system itself that confers noise tolerance in shallow mRNA-seq measurements. Related work in neuroscience has explored the implications of hierarchical network architecture for learning the dominant dimensions of data (Saxe et al., 2013; Hinton and Salakhutdinov, 2006).

Discovering and exploiting low dimensionality to reduce uncertainty in measurements is at the heart of modern signal processing techniques (Donoho 2006; Candès et al., 2006). These methods first found success in imaging applications, where low dimensionality arises from the statistics and redundancies of natural images, enabling most images to be accurately represented by a small number of wavelets or other basis functions. Our results suggest that shallow mRNA-seq is similarly enabled by an inherent low dimensionality in gene expression datasets that emerges from groups of covarying genes. Just as only a few wavelets are needed to represent most images, only a few groups of transcriptional programs seem to be necessary to produce a coarse-grained representation of transcriptional state.

We believe that the measurement of many diverse biological systems could benefit from the identification and analysis of hidden low-dimensional representations. For instance, proteome quantification, protein-protein interactions, and human genetic variant data all contain high levels of correlations, suggesting these datasets may all be effectively low dimensional. We anticipate new modes of biological inquiry as advances from signal processing are integrated into biological data analysis and as the underlying structural features of biological networks are exploited for large-scale measurements.

EXPERIMENTAL PROCEDURES

Simulated Shallow Sequencing through Down-sampling of Reads

Transcriptional datasets were obtained from the GEO (Zeisel et al. [2015] was from http://www.linnarssonlab.org). mRNA-seq read counts were normalized by the total number of reads in the sample. For each read depth, we model the sequencing noise with a multinomial distribution. The Zeisel et al. (2015) data were sampled without replacement because of the unique molecular identifiers (see Supplemental Experimental Procedures).

Finding Genes Significantly Associated with a Principal Component

We first generated a null distribution of gene loadings from the principal components of a shuffled, transcript-count matrix. All p values were computed with respect to this distribution; averages over 15 replicates are reported.

Gene Set Enrichment Analysis

GSEA was performed with 1,370 gene lists from MSigDB (Subramanian et al., 2005). The loadings of each principal component were collected in a distribution and loadings within 2 SDs from the mean of this distribution were considered for analysis. We applied a hypergeometric test with a significance p value cutoff of 10^{-4}.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures and five figures and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2016.04.001.

AUTHOR CONTRIBUTIONS

G.H., H.E.-S., and M.T. conceived the idea. G.H. wrote the simulations and analyzed data, with input from M.T. and H.E.-S. R.B. and M.T. performed theoretical analysis. R.B. wrote the mathematical proofs. The manuscript was written by G.H., R.B., H.E.-S., and M.T.

ACKNOWLEDGMENTS

The authors would like to thank Jason Kreisberg, Alex Fields, David Sivak, Patrick Cahan, Jonathan Weissman, Chun Ye, Michael Chevalier, Satwik Rajaram, and Steve Altschuler for careful reading of the manuscript; Eric Chow, John Haliburton, Sisi Chen, and Emeric Charles for their experimental insights; and Paul Rivaud for website design assistance. This work was supported by the UCSF Center for Systems and Synthetic Biology (NIGMS P50 GM081879). H.E.S. acknowledges support from the Paul G. Allen Family Foundation. M.T. acknowledges support from the NIH Office of the Director, the National Cancer Institute, and the National Institute of Dental and Craniofacial Research (NIH DP5 OD012194).

Received: November 30, 2015
Revised: March 8, 2016
Accepted: April 4, 2016
Published: April 27, 2016

REFERENCES

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., and Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA *96*, 6745–6750.

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

Bengio, Y., Delalleau, O., Le Roux, N., Paiement, J.-F., Vincent, P., and Ouimet, M. (2004). Learning eigenfunctions links spectral embedding and kernel PCA. Neural Comput. *16*, 2197–2219.

Bergmann, S., Ihmels, J., and Barkai, N. (2003). Iterative signature algorithm for the analysis of large-scale gene expression data. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. *67*, 031902.

Bonneau, R. (2008). Learning biological networks: from modules to dynamics. Nat. Chem. Biol. *4*, 658–664.

Candès, E.J., Romberg, J.K., and Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. *59*, 1207–1223.

Ding, C., and He, X. (2004). K-means clustering via principal component analysis. ICML Proceedings of the 21st International Conference on Machine Learning (ACM), p. 29.

Donoho, D.L. (2006). Compressed sensing. IEEE Trans. Inf. Theory *52*, 1289–1306.

Duarte, M.F., Davenport, M.A., Takbar, D., Laska, J.N., Sun, T., Kelly, K.F., and Baraniuk, R.G. (2008). Single-pixel imaging via compressive sampling. IEEE Signal Process. Mag. *25*, 83–91.

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

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA *95*, 14863–14868.

Fan, H.C., Fu, G.K., and Fodor, S.P.A. (2015). Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. Science *347*, 1258367.

Ham, J., Lee, D.D., Mika, S., and Schölkopf, B. (2004). A kernel view of the dimensionality reduction of manifolds. ICML Proceedings of the 21st International Conference on Machine Learning (ACM), p. 47.

Hinton, G.E., and Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. Science *313*, 504–507.

Holter, N.S., Maritan, A., Cieplak, M., Fedoroff, N.V., and Banavar, J.R. (2001). Dynamic modeling of gene expression data. Proc. Natl. Acad. Sci. USA *98*, 1693–1698.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., and Amit, I. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science *343*, 776–779.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.

Kliebenstein, D.J. (2012). Exploring the shallow end; estimating information content in transcriptomics studies. Front. Plant Sci. *3*, 213.

Kumar, R.M., Cahan, P., Shalek, A.K., Satija, R., DaleyKeyser, A.J., Li, H., Zhang, J., Pardee, K., Gennert, D., Trombetta, J.J., et al. (2014). Deconstructing transcriptional heterogeneity in pluripotent stem cells. Nature *516*, 56–61.

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.

Ng, A.Y., Jordan, M.I., and Weiss, Y. (2001). On spectral clustering: analysis and an algorithm. In Advances in Neural Information Processing Systems (MIT Press), pp. 849–856.

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science *344*, 1396–1401.

Pollen, A.A., Nowakowski, T.J., Shuga, J., Wang, X., Leyrat, A.A., Lui, J.H., Li, N., Szpankowski, L., Fowler, B., Chen, P., et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nat. Biotechnol. *32*, 1053–1058.

Ringnér, M. (2008). What is principal component analysis? Nat. Biotechnol. *26*, 303–304.

Roweis, S.T., and Saul, L.K. (2000). Nonlinear dimensionality reduction by locally linear embedding. Science *290*, 2323–2326.

Saxe, A.M., Mcclelland, J.L., and Ganguli, S. (2013). Learning hierarchical category structure in deep neural networks. Proceedings of the 35th Annual Meeting of the Cognitive Science Society, pp. 1271–1276.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet. *34*, 166–176.

Shai, R., Shi, T., Kremen, T.J., Horvath, S., Liau, L.M., Cloughesy, T.F., Mischel, P.S., and Nelson, S.F. (2003). Gene expression profiling identifies molecular subtypes of gliomas. Oncogene *22*, 4918–4923.

Shalek, A.K., Satija, R., Adiconis, X., Gertner, R.S., Gaublomme, J.T., Raychowdhury, R., Schwartz, S., Yosef, N., Malboeuf, C., Lu, D., et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature *498*, 236–240, advance online publication.

Shalek, A.K., Satija, R., Shuga, J., Trombetta, J.J., Gennert, D., Lu, D., Chen, P., Gertner, R.S., Gaublomme, J.T., Yosef, N., et al. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature *510*, 363–369.

Shankar, R. (2012). Principles of Quantum Mechanics (Springer Science & Business Media).

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V., and Ren, B. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116–120.

Stewart, G.W., and Sun, J. (1990). Matrix Perturbation Theory (Academic Press).

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Res. *25*, 1491–1498.

Treutlein, B., Brownfield, D.G., Wu, A.R., Neff, N.F., Mantalas, G.L., Espinoza, F.H., Desai, T.J., Krasnow, M.A., and Quake, S.R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature *509*, 271–375.

Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. *9*, 85.

Velculescu, V.E., Madden, S.L., Zhang, L., Lash, A.E., Yu, J., Rago, C., Lal, A., Wang, C.J., Beaudry, G.A., Ciriello, K.M., et al. (1999). Analysis of human transcriptomes. Nat. Genet. *23*, 387–388.

Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science *347*, 1138–1142.