SUPPLEMENTAL INFORMATION
Table of Contents

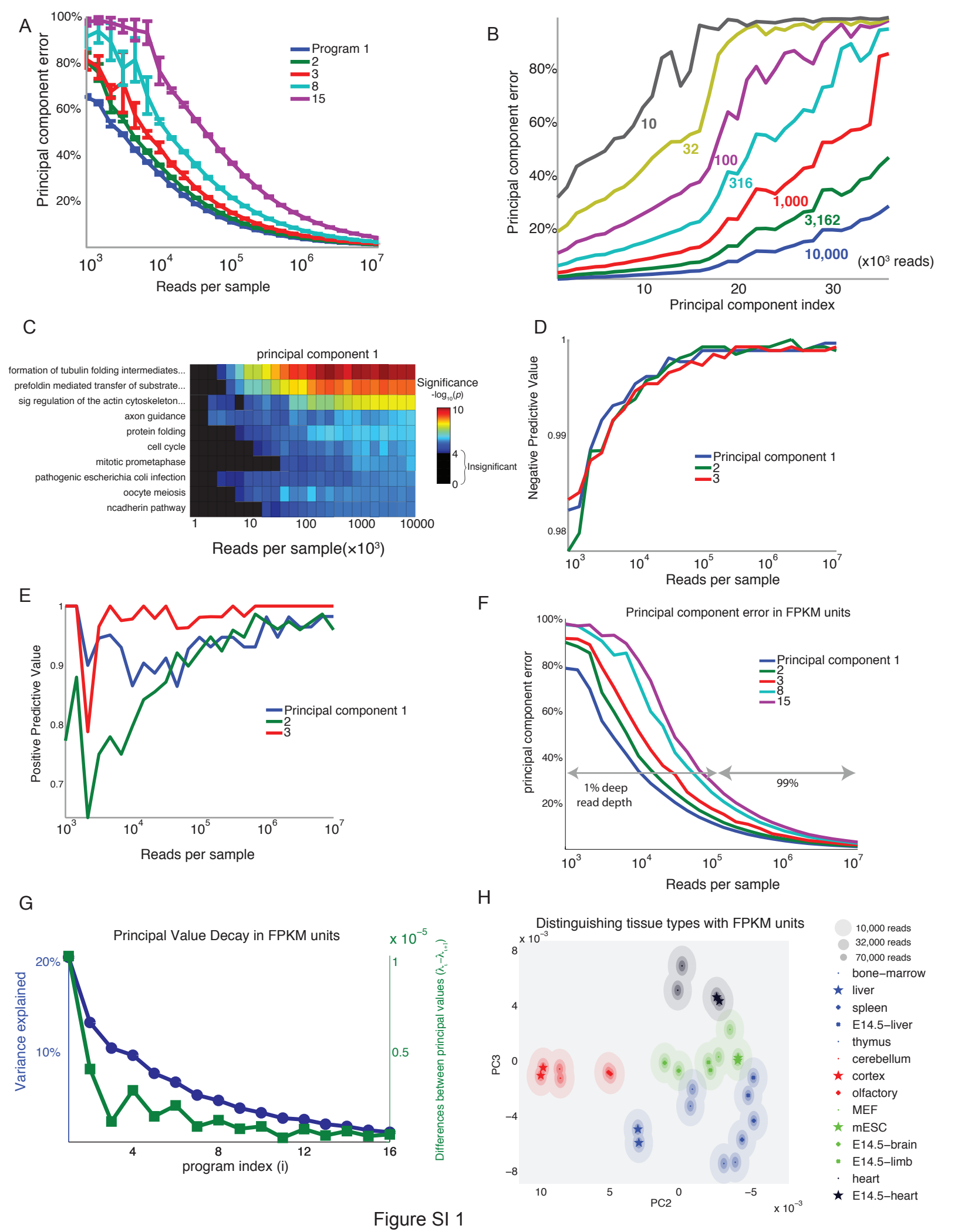

A
Principal component error
100%
80%
60%
40%
20%
Program 1
2
3
8
15
10^3
10^4
10^5
10^6
10^7
Reads per sample
B
Principal component error
80%
60%
40%
20%
10
32
100
316
1,000
3,162
10,000
(x10³ reads)
10
20
30
Principal component index
C
principal component 1
formation of tubulin folding intermediates...
prefoldin mediated transfer of substrate...
sig regulation of the actin cytoskeleton...
axon guidance
protein folding
cell cycle
mitotic prometaphase
pathogenic escherichia coli infection
oocyte meiosis
ncadherin pathway
Significance
$-\log_{10}(p)$
10
4
0
Insignificant
1
10
100
1000
10000
Reads per sample(×10³)
D
Negative Predictive Value
1
0.99
0.98
Principal component 1
2
3
10^3
10^4
10^5
10^6
10^7
Reads per sample
E
Positive Predictive Value
1
0.9
0.8
0.7
Principal component 1
2
3
10^3
10^4
10^5
10^6
10^7
Reads per sample
F
Principal component error in FPKM units
principal component error
100%
80%
60%
40%
20%
Principal component 1
2
3
8
15
1% deep
read depth
99%
10^3
10^4
10^5
10^6
10^7
Reads per sample
G
Principal Value Decay in FPKM units
Variance explained
20%
10%
x 10^{-5}
1
0.5
Differences between principal values ($\lambda_i - \lambda_{i+1}$)
4
8
12
16
program index (i)
H
Distinguishing tissue types with FPKM units
x 10^{-3}
8
4
0
−4
−8
PC3
10
5
0
−5
PC2
x 10^{-3}
10,000 reads
32,000 reads
70,000 reads
bone−marrow
liver
spleen
E14.5−liver
thymus
cerebellum
cortex
olfactory
MEF
mESC
E14.5−brain
E14.5−limb
heart
E14.5−heart

Figure SI 1

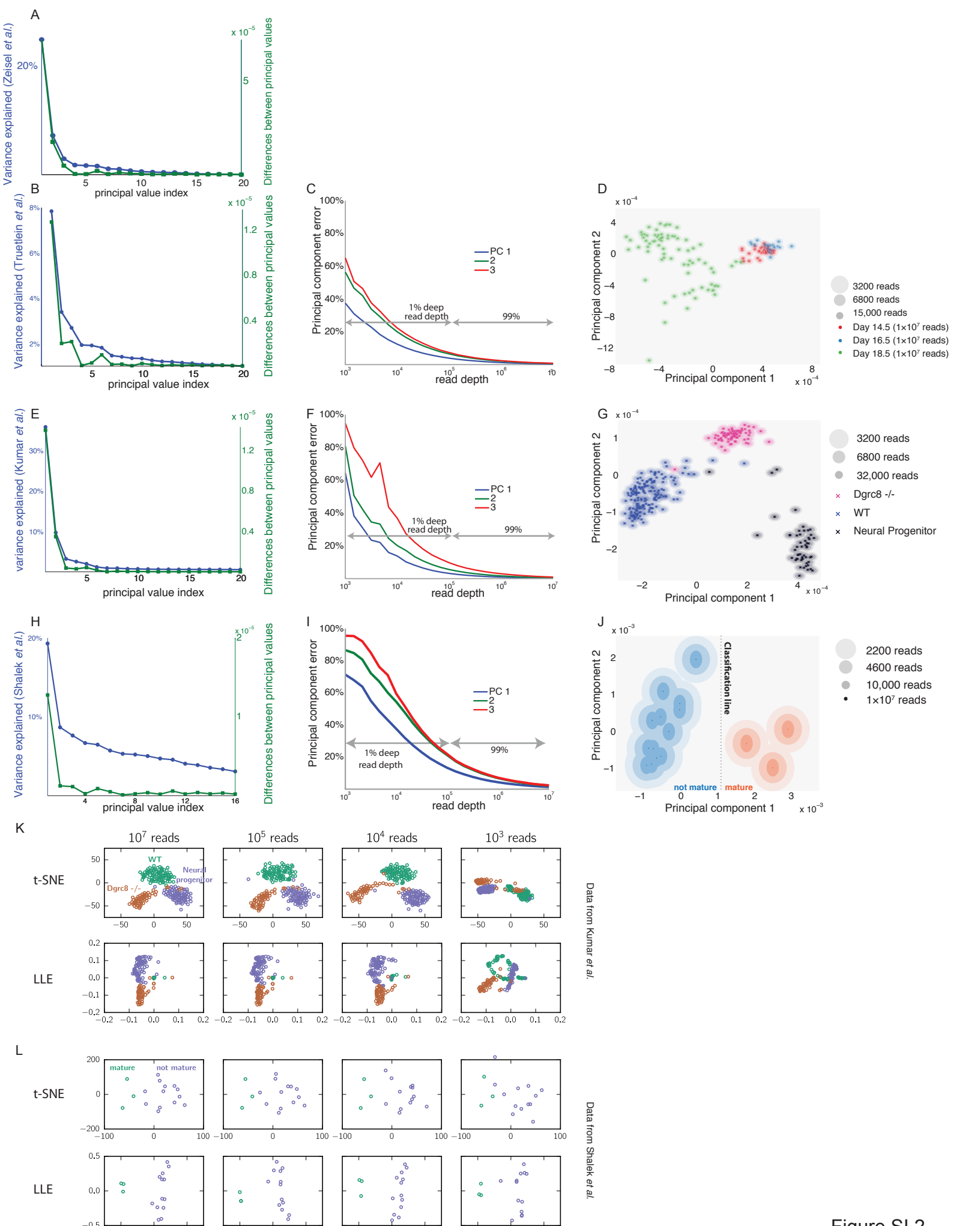

A
Variance explained (Zeisel et al.)
Differences between principal values
principal value index
B
Variance explained (Tuetlein et al.)
principal value index
C
Principal component error
PC 1
2
3
1% deep read depth
99%
read depth
D
Principal component 2
Principal component 1
3200 reads
6800 reads
15,000 reads
Day 14.5 (1×10⁷ reads)
Day 16.5 (1×10⁷ reads)
Day 18.5 (1×10⁷ reads)
E
variance explained (Kumar et al.)
principal value index
F
Principal component error
read depth
G
Principal component 2
Principal component 1
3200 reads
6800 reads
32,000 reads
Dgrc8 -/-
WT
Neural Progenitor
H
Variance explained (Shalek et al.)
principal value index
I
Principal component error
read depth
J
Principal component 2
Principal component 1
Classification line
not mature
mature
2200 reads
4600 reads
10,000 reads
1×10⁷ reads
K
10⁷ reads
10⁵ reads
10⁴ reads
10³ reads
t-SNE
LLE
WT
Neural progenitor
Dgrc8 -/-
Data from Kumar et al.
L
t-SNE
LLE
mature
not mature
Data from Shalek et al.

Figure SI 2

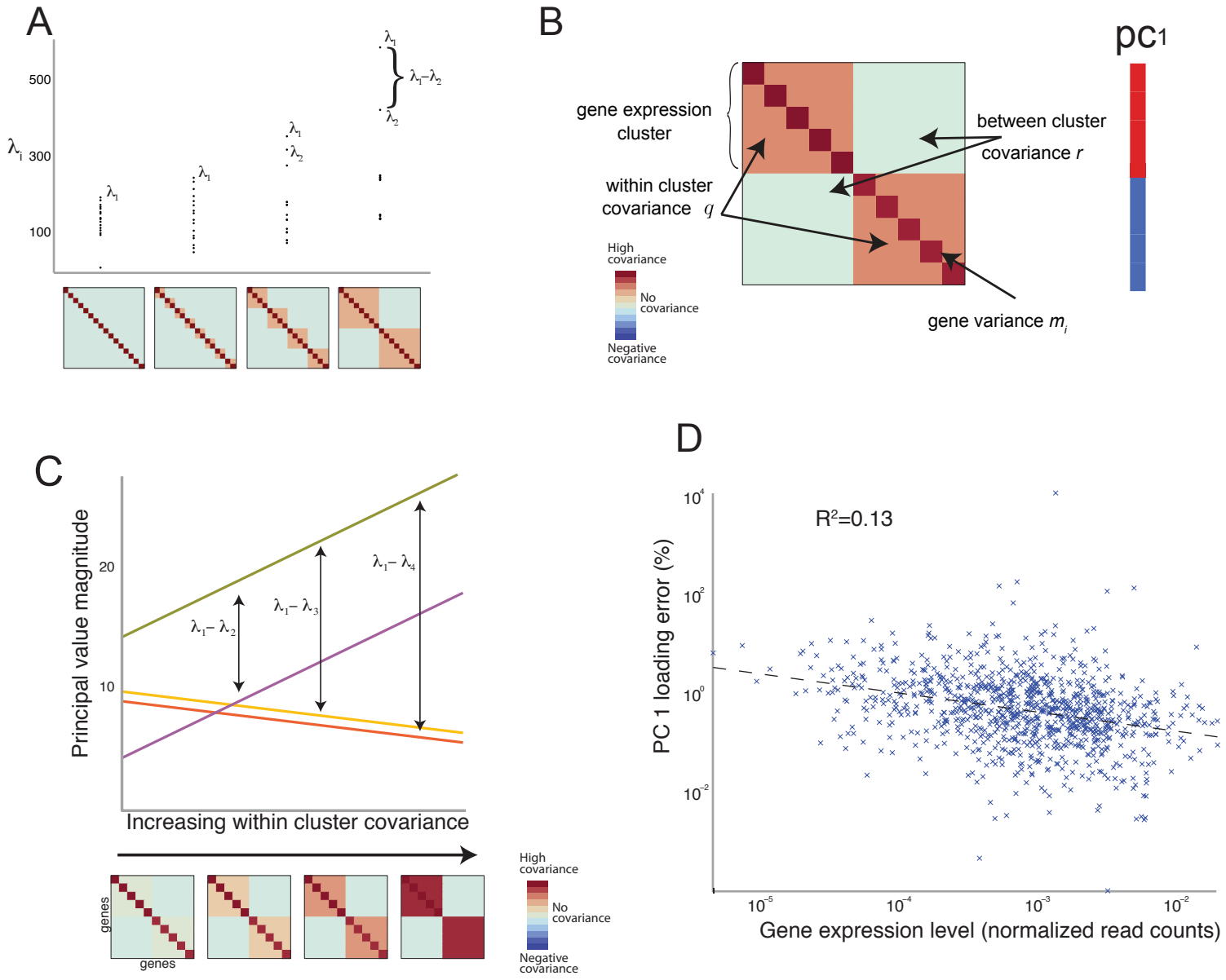
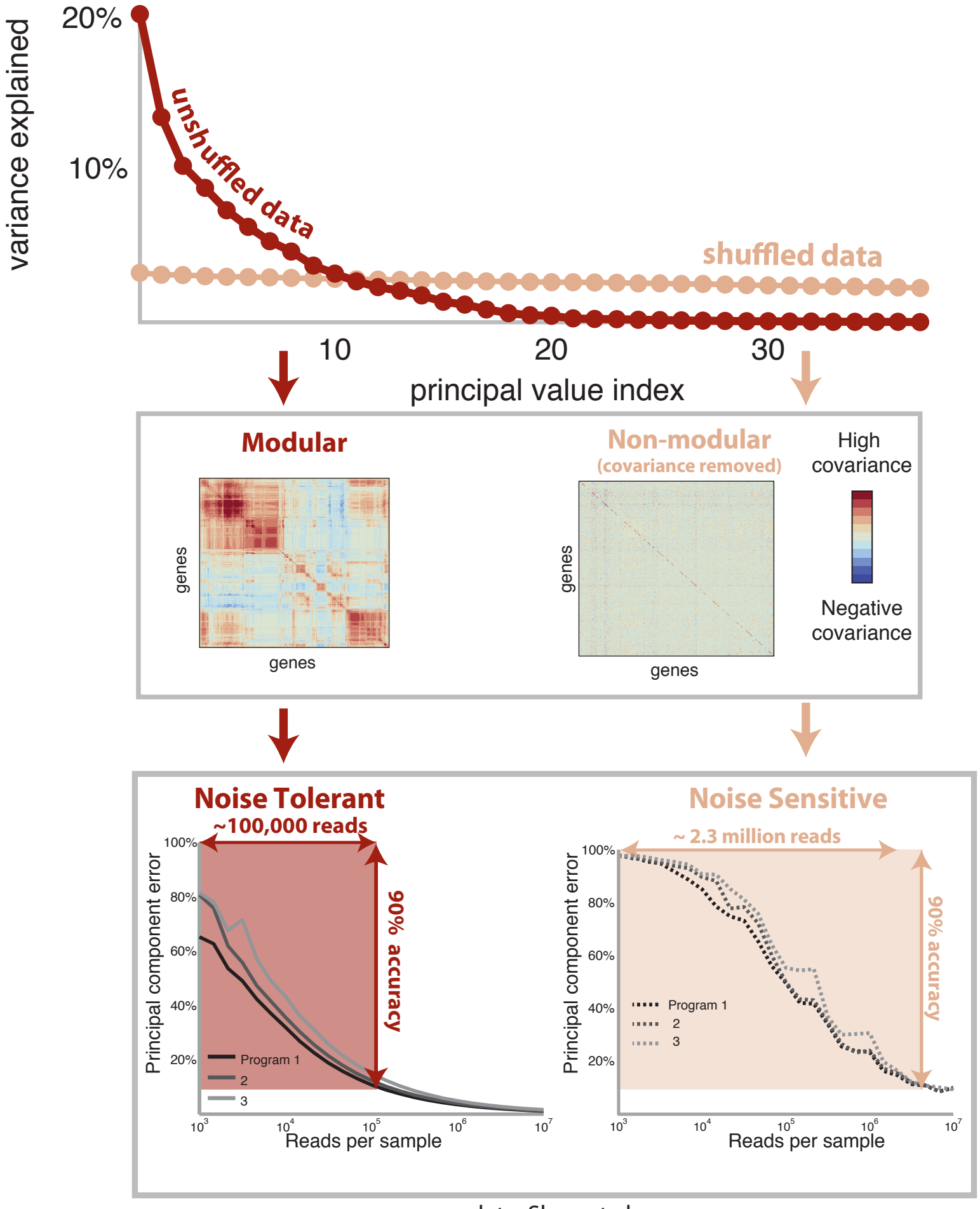

A
500
300
100
λ_1
λ_1–λ_2
B
pc1
gene expression cluster
between cluster covariance r
within cluster covariance q
gene variance m_i
High covariance
No covariance
Negative covariance
C
Principal value magnitude
20
10
λ_1– λ_2
λ_1– λ_3
λ_1– λ_4
Increasing within cluster covariance
genes
genes
High covariance
No covariance
Negative covariance
D
R^2=0.13
PC 1 loading error (%)
10^4
10^2
10^0
10^{-2}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
Gene expression level (normalized read counts)

Figure SI 3



data: Shen et al.

Figure SI 4

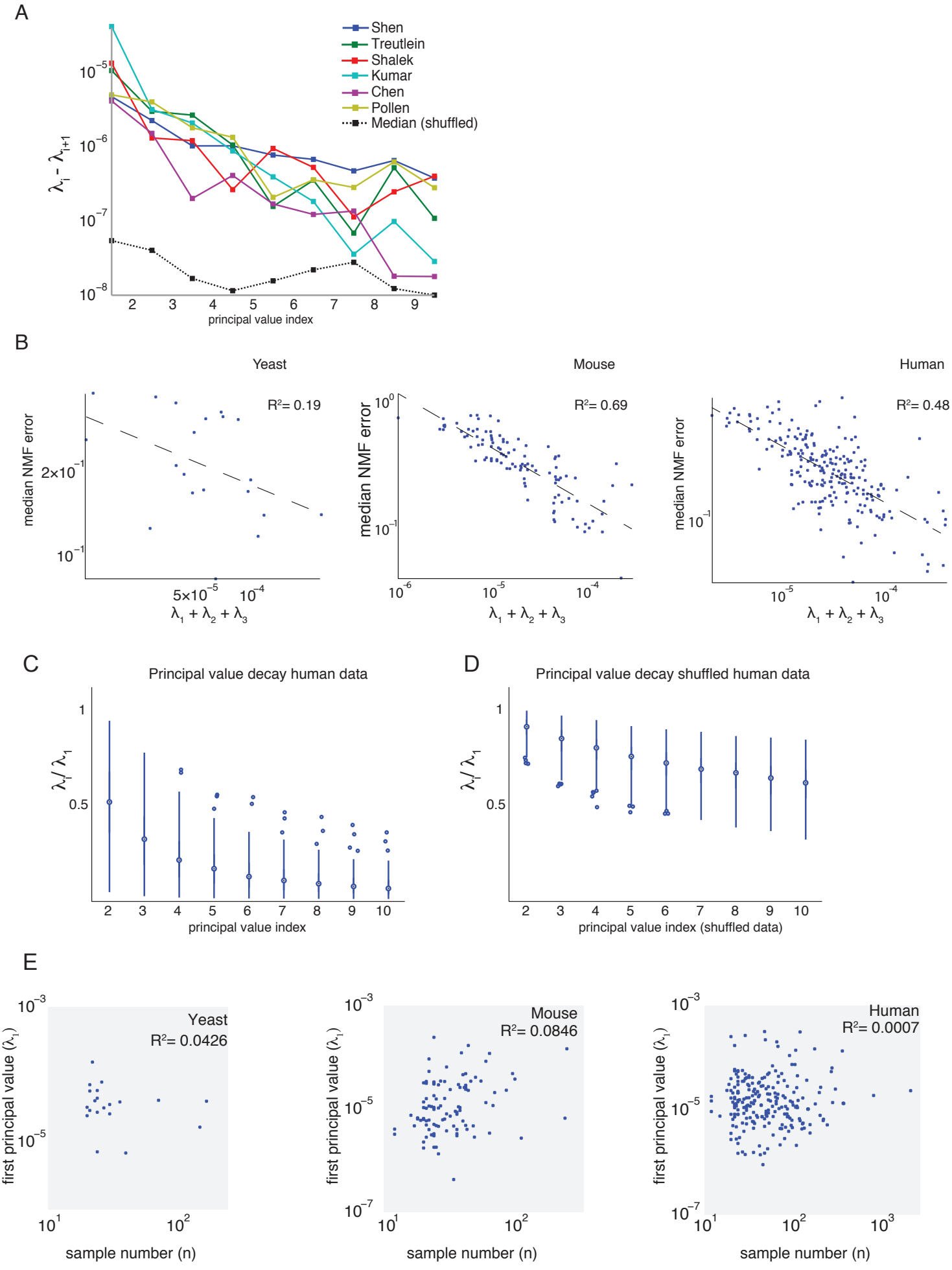

A
Shen
Treutlein
Shalek
Kumar
Chen
Pollen
Median (shuffled)
principal value index
B
Yeast
Mouse
Human
R²= 0.19
R²= 0.69
R²= 0.48
median NMF error
C
Principal value decay human data
principal value index
D
Principal value decay shuffled human data
principal value index (shuffled data)
E
Yeast
R²= 0.0426
Mouse
R²= 0.0846
Human
R²= 0.0007
first principal value (λ1)
sample number (n)

Figure SI 5

1 Supplemental Figure Legends

Figure S1: Stability of principal component error and gene set enrichment analysis across down-sampling replicates for the Shen et al. dataset, related to Figure 2
(A) Mean (solid lines) and standard deviation (error bars) in principal component error of mouse tissue data (Shen et al.), as a function of read depth as calculated from 20 simulated shallow sequencing experiments at each of 25 indicated read depths. Narrow width of error bars illustrates the stability of the PCA error calculation to the downsampling procedure. The mean PCA error curves are also shown in Figure 2A.
(B) Principal component error for first 38 principal components of the mouse tissue data at 7 read depths, illustrating the number of principal components that can be accurately reconstructed as sample read depth is decreased. For example, nine principal components can be reconstructed at less than 20% error with only 133,000 reads per sample.
(C) Gene Set Enrichment Analysis for principal component 1 of the mouse tissue dataset at decreasing read depth. Significant gene sets (see scale bar) are stable even below $32,000$ reads. Figure 2 focuses on analysis of principal components 2 and 3 as they are of more biological relevance for classification.
(D) Negative Predictive Value of Gene Set Enrichment Analysis applied to mouse tissue data for the first three principal components (color) over a large range of read depths. Negative Predictive Value indicates the fraction of gene sets correctly considered insignificant out of all gene sets considered insignificant.
(E) Positive Predictive Value of Gene Set Enrichment Analysis applied to mouse tissue data for the first three principal components (color) over a large range of read depths. Positive Predictive Value indicates the fraction of gene sets correctly considered statistically significant out of all gene sets considered statistically significant.
(F) Principal component error as a function of read depth for selected principal components for the Shen et al. data as in Figure 2A. Here the transcriptional programs are calculated from the FPKM values rather then read count data. Again the first three principal components can be recovered with >80% accuracy with just 1% of the traditional read depths. 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.
(G) Variance explained by transcriptional program (blue) and differences between principal values (green) calculated from the FPKM values. Like the read count 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 sequencing noise.
(H) Projection of a subset of the mouse tissue data onto principal components two and three as in Figure 2. Here, principal components are calculated with FPKM values, rather than read count data. As in Figure 2D, the ellipses represent uncertainty due to sequencing noise at specific reads depths. Again, similar tissues lie close together. Transcriptional program two separates neural tissues from non-neural tissues while transcriptional program three distinguishes tissues involved in haematopoiesis from other tissues.

Figure S2: Principal value separation is large in single cell mRNA-seq datasets, related to Figure 3

(A) Variance explained by principal components (blue) and differences between principal values (green) of the Zeisel et al. data. Similar to the bulk mRNA-seq 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. See Figure 3 for the principal component error and cell-type classification accuracy as a function of transcript coverage.

(B) Variance explained by principal components (blue) and differences between principal values (green) of the Treutlein et al. data.

(C) Principal component error as a function of read depth for the first three principal components for the Treutlein et al. data.

(D) Transcriptional state of single cells during lung development at two time points E16.5 and 18.5 from Treutlein et al. projected onto the first two principal components (see Supplemental Experimental Procedures). Radii indicate error at given read depth. Developmental stages corresponding to nascent (16.5) and mature (18.5) progenitor cells can be distinguished at 3,200 reads.

(E) Variance explained by principal components (blue) and differences between principal values (green) of the Kumar et al. data.

(F) Principal component error as a function of read depth for the first three principal components for the Kumar et al. data.

(G) Projection of the transcriptional state of wild type and Dgcr8 knockout mouse embryonic stem cells from Kumar et al. on the first two principal components. The wild type cells separate from the knockout cells which are deficient in miRNA processing at 3,200 reads.

(H) Variance explained by principal components (blue) and differences between principal values (green) of the Shalek et al. data.

(I) Principal component error as a function of read depth for the first three principal components for the Shalek et al. data.

(J) Projection of the transcriptional state of 18 bone-marrow-derived dendritic single cells from Shalek et al. data on the first two principal components. The "mature" and "not mature" cells are distinguishable at 3,200 reads.

(K) and (L) Distinct transcriptional states in single cells can be uncovered by the nonlinear unsupervised learning methods t-SNE and LLE at low read depth. Computed clusters in Kumar et al. data and Shalek et al. data at 10^5 reads are almost identical to those obtained at 10^7 reads, with significant information preserved even at 10^3 reads.

Figure S3: Impact of gene expression covariance and absolute gene expression level on principal value separation, related to Figure 4

(A) Principal value separation increases with module size. Principal values λ_i shown for a sixteen gene system for increasing module size b (with $q = 40$, $r = -8$, and m_i constant within blocks and spanning $[80, 200]$).

(B) Block-diagonal covariance matrix for a general model of gene expression analyzed in the Supplemental Information Section 2.2 (the ten gene, two module case is illustrated). The matrix is annotated with relevant parameters, q, r, m_i, b. The first principal component discriminates membership in the two underlying gene expression modules.

(C) Principal value separation increases or remains constant as within-cluster covariance q increases. Clustered covariance matrices depicted along x-axis and the first four principal values are analytically determined from the gene expression model of (B) as described in

Supplemental Information Section 2.2, with $b = 4,\ r = -1,\ m_1 = 10,\ m_2 = 6$.
(D) Principal component loading error versus absolute gene expression level for the Shen et al. dataset. For each gene, the absolute gene expression level is normalized read counts summed across samples. The gene-wise loading error is calculated for gene i as $|\mathbf{pc}_{1,i} - \hat{\mathbf{pc}}_{1,i}|/\mathbf{pc}_{1,i}$ at a read depth of 46,000 reads per sample. The weak correlation ($R^2 = 0.13$) indicates that absolute gene expression level does not significantly contribute to the gene-wise principal component error.

Figure S4: The modularity of gene expression enables accurate, low depth transcriptional program identification in single cell mRNA-seq data, related to Figure 4
The gene expression covariance matrix of the Shen et al. data reveals large modules of covarying genes (middle), whose signature is a few, dominant transcriptional programs that explain relatively large variances in the data (top). As predicted by the model, these dominant transcriptional programs are robust to low-coverage profiling (bottom). Shuffling the Shen et al. data destroys the modular structure, resulting in noise-sensitive transcriptional programs. For the shuffled data, $\sim$2.3 million reads are required for 90% accuracy of the first three transcriptional programs, whereas $\sim 100,000$ reads suffices for the original dataset.

Figure S5: Principal value separation is common in mRNA-seq and microarray datasets and is due to gene expression covariance, related to Figure 5
(A) Principal value separations $\lambda_i - \lambda_{i+1}$ for six mRNA-seq datasets illustrating the generality of principal value decay. Median of the differences between principal values for the datasets after shuffling is shown in black.
(B) Relationship between error in Non-negative Matrix Factorization (NMF) and sum of first three principal values of microarray data. NMF error calculated as median error across three NMF parts at 45,000 reads per sample (see Supplemental Experimental Procedures). As the summed principal values on the x-axis represent the variance explained by the first three principal components, this indicates that the performance of NMF correlates with the existence of dominant, transcriptional programs.
(C) Ratio of principal values, λ_i/λ_1, in human microarray data for $2 \leq i \leq 10$. Principal values are well-separated.
(D) After shuffling datasets to remove gene expression covariance, the principal values are no longer well-separated.
(E) Sample number (x-axis) and the magnitude of the first principal value λ_1 (y-axis) are approximately uncorrelated. Left: 20 yeast datasets, $R^2 = 0.04$. Middle: 106 mouse datasets, $R^2 = 0.08$. Right: 226 human datasets, $R^2 = 7.0 \times 10^{-4}$.

2 Supplemental Theory

2.1 Shallow sequencing

This section develops the theoretical framework used in the main text to analyze shallow sequencing. We use perturbation theory to find how principal components of shallow data differ from those of deep data and explore this relationship in the context of a simple, multinomial noise model for mRNA-sequencing. In the process, we provide background on Equation (1) and derive Equation (2) of the main text.

We begin by summarizing and extending the notation of the main text.

2.1.1 Introduction and Notation

Suppose we have collected reads from deep mRNA-seq experiments in a matrix $\boldsymbol{G}$ of dimensions $g \times n$, where g is the number of genes in the genome and n is the number of experimental samples analyzed. Each entry satisfies $0 \leq G_{ij} \leq N_{\text{deep}}$, where N_{deep} is the total number of reads collected in each sample and is assumed (for convenience) to be constant across samples. Then P^0_{ij}, the probability that a transcript from sample j maps to gene i, is equal to G_{ij}/N_{deep}. We assume that N_{deep} is large enough that P^0_{ij} represents the true, underlying probabilities of gene expression. It is frequently more convenient to work with the row-centered probabilities, $P_{ij} \triangleq P^0_{ij} - n^{-1}\sum_j P^0_{ij}$. For instance, the *deep gene covariance matrix* $\boldsymbol{C}$, of dimensions $g \times g$, is of fundamental interest and can be written simply as $\boldsymbol{C} \triangleq \boldsymbol{P}\boldsymbol{P}^{\mathsf{T}}/(n-1)$, where $\boldsymbol{P}$ is a matrix with entries P_{ij}. Note that while we define the covariance matrix in terms of transcript probabilities, we could similarly define the covariance matrix in terms of transcripts measured in FPKM units, by rescaling each P_{ij} by a gene length dependent factor. We choose to work with transcript probabilities for mathematical convenience.

Now assume that we repeat the sequencing experiments with only $N \ll N_{\text{deep}}$ reads. From these shallow mRNA-seq experiments, we obtain data $\hat{G}_{ij}$ from which we compute the gene expression probabilities $\hat{P}_{ij}$ and gene-gene covariances $\hat{C}_{ij}$, which we collect in matrices $\hat{\boldsymbol{P}}$ and $\hat{\boldsymbol{C}}$. (In general, we put hats on the quantities calculated from shallow data.) We are primarily interested in minimizing the number of reads while preserving the biologically relevant information contained in $\hat{\boldsymbol{C}}$. In particular, this section addresses the question of *how does sequencing depth N affect the distance between the i^{th} principal component of $\hat{\boldsymbol{P}}$ and the i^{th} principal component of $\boldsymbol{P}$?*

The *principal components* $\boldsymbol{v}_i$ and *principal values* λ_i of the probability matrix $\boldsymbol{P}$ are the eigenvectors and eigenvalues of the covariance matrix $\boldsymbol{C}$ and therefore satisfy

$$\boldsymbol{C}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i. \tag{2.1a}$$

We adopt the convention that the eigenvectors are sorted by decreasing eigenvalue, so $\boldsymbol{v}_1$ is the first eigenvector of $\boldsymbol{C}$, corresponding to the direction of maximum variance in the data $\boldsymbol{P}$. Similarly, the *shallow principal components* and *shallow principal values* satisfy

$$\hat{\boldsymbol{C}}\hat{\boldsymbol{v}}_i = \hat{\lambda}_i \hat{\boldsymbol{v}}_i. \tag{2.1b}$$

The rest of this section is structured as follows. Section 2.1.2 describes a simple, multinomial noise model for mRNA-seq and describes how noise propagates to the shallow covariance matrix. Section 2.1.3 explains how this noise "perturbs" the deep covariance matrix and bounds the resulting change in the first principal component, thereby deriving Equation (2) of the main text. Finally, Section 2.1.4 generalizes this result to higher principal components.

We additionally use the following notation. A matrix $\boldsymbol{X}$ has elements X_{ij}. The transpose of a vector $\boldsymbol{x}$ is $\boldsymbol{x}^\mathsf{T}$. Expectation is denoted $\mathbb{E}\cdot$ and variance by $\mathbb{V}\cdot$. We use $\|\cdot\|$ or $\|\cdot\|_2$ to mean the ℓ^2 norm of a vector and the spectral norm of a matrix (*i.e.* the maximum singular value of the matrix). We write $\|\cdot\|_1$ for the ℓ_1 norm and $\|\cdot\|_\infty$ for the infinity norm of a matrix (*i.e.* the maximum absolute row sum).

2.1.2 Noise Model

We first introduce a noise model that describes the impact of sequencing depth on the data $\boldsymbol{G}$. While there are many sources of noise in the measurement of mRNA-transcripts, we begin with a simplifying assumption.

Assumption 1. *The dominant source of noise in the measurement of mRNA-transcripts is counting noise. Further, all noise, across both genes and samples, is uncorrelated.*

Hence the data collected from shallow sequencing is

$$\hat{G}_{ij} \sim \text{Binomial}(N, P^0_{ij}).$$

The maximum likelihood estimate of the true (*i.e.* obtained from deep data) gene expression probabilities is simply

$$\hat{\boldsymbol{P}}^{\boldsymbol{0}} = \hat{\boldsymbol{G}}/N.$$

Using the assumption that all noise (across both genes and samples) is uncorrelated and further assuming that the binomial is well-approximated by the normal distribution, we have that the underlying probabilities are

$$\hat{P}^0_{ij} \sim \text{Normal}\left(P^0_{ij}, \frac{1}{N}P^0_{ij}(1-P^0_{ij})\right)$$

which, when row-centered, are

$$\hat{P}_{ij} \sim \text{Normal}\left(P^0_{ij} - \frac{1}{n}\sum_j P^0_{ij},\ \ \frac{1}{N}P^0_{ij}(1-P^0_{ij})\right). \tag{2.2}$$

With this notation, the *shallow gene covariance* is $\hat{\boldsymbol{C}} \triangleq \hat{\boldsymbol{P}}\hat{\boldsymbol{P}}^\mathsf{T}/(n-1)$.

Similar models for sequencing noise, as well as more specialized models for single-cell RNA-seq, have been recently proposed (Marioni et al. 2008; McIntyre et al. 2011; Pollen et al. 2014; Liu, Zhou, and White 2014; Tarazona et al. 2011; Anders and Huber 2010; Islam et al. 2014; Shiroguchi et al. 2012; Grün, Kester, and van Oudenaarden 2014; Brennecke et al. 2013; Ding et al. 2015; Daley and Smith 2014; Vallejos, Marioni, and Richardson 2015). Our goal in what follows is to use the simplest noise model to identify the important parameters and capture their basic dependencies. However, more realistic noise models will fit comfortably in our framework.

To measure the error induced by shallow sequencing, we introduce two definitions.

Definition 1. *The* error in gene expression probabilities *is* $\boldsymbol{E} \triangleq \hat{\boldsymbol{P}} - \boldsymbol{P}$.

From equation (2.2), this error is distributed as

$$E_{ij} \sim \text{Normal}\left(0, \frac{1}{N} P_{ij}^0 (1 - P_{ij}^0)\right). \tag{2.3}$$

Definition 2. *The* covariance distortion *induced by shallow sequencing is* $\boldsymbol{D} \triangleq \hat{\boldsymbol{C}} - \boldsymbol{C}$.

With the definitions of gene covariance as well as Definition 1, the covariance distortion can be expanded as

$$\boldsymbol{D} = \frac{1}{n-1}(\boldsymbol{P}\boldsymbol{E}^\mathsf{T} + \boldsymbol{E}\boldsymbol{P}^\mathsf{T} + \boldsymbol{E}\boldsymbol{E}^\mathsf{T}). \tag{2.4}$$

2.1.3 Perturbation theory

Our goal is to find how the principal components of $\boldsymbol{P}$ differ from those of $\hat{\boldsymbol{P}}$. Our approach treats the covariance distortion $\boldsymbol{D}$ as a (random) perturbation to the deep covariance matrix $\boldsymbol{C}$. We then use a result from perturbation theory as well as the properties of the noise model of Section 2.1.2 to find the resulting change in the principal components of $\boldsymbol{P}$. Along the way, we introduce assumptions that reflect properties of biological data that are needed to simplify the result.

Our main tool from perturbation theory (Stewart and Sun 1990; Shankar 2012) describes how the eigenvectors of a positive semi-definite matrix change when the matrix is perturbed:

Proposition 1. *Let* $\boldsymbol{C}$ *be a positive semi-definite matrix with eigenvalues* λ_k^0 *and eigenvectors* v_k^0. *Further let*

$$\hat{\boldsymbol{C}}(\epsilon) = \boldsymbol{C} + \epsilon \boldsymbol{D}$$

be a perturbation of $\boldsymbol{C}$. *With some weak assumptions on* $\boldsymbol{D}$, *the eigenvalues and eigenvectors of* $\hat{\boldsymbol{C}}$ *are*

$$\hat{\lambda}_k(\epsilon) = \lambda_k^0 + \epsilon \lambda_k^1$$
$$\hat{\boldsymbol{v}}_k(\epsilon) = \boldsymbol{v}_k^0 + \epsilon \boldsymbol{v}_k^1$$

where the first-order corrections to the eigenvalues and eigenvectors are

$$\lambda_k^1 = \boldsymbol{v}_k^{0\mathsf{T}} \boldsymbol{D} \boldsymbol{v}_k^0$$

and

$$\boldsymbol{v}_k^1 = \sum_{j \neq k} \frac{\boldsymbol{v}_j^{0\mathsf{T}} \boldsymbol{D}\, \boldsymbol{v}_k^0}{\lambda_k^0 - \lambda_j^0} \boldsymbol{v}_j^0 + a_k \boldsymbol{v}_k^0.$$

Here a_k *is a constant that is determined by the constraint that* $\hat{\boldsymbol{v}}_k$ *is unit length.*

Equation (1) of the main text immediately follows from this proposition. As explained in the main results, a natural measure of the error induced by shallow sequencing in the kth

principal component is $\|\hat{\boldsymbol{v}}_k - \boldsymbol{v}_k\|_2$. From Proposition 1, this quantity is to first order

$$\|\hat{\boldsymbol{v}}_k - \boldsymbol{v}_k\|_2 \approx \sqrt{a_k^2 + \sum_{j \neq k} \left(\frac{\boldsymbol{v}_j^\mathsf{T} \boldsymbol{D}\, \boldsymbol{v}_k}{\lambda_k - \lambda_j} \right)^2}. \qquad \text{(main text equation 1)}$$

In this formula, a_k can be determined by the convention that $\hat{\boldsymbol{v}}_k$ has unit length,

$$(1 + a_k)^2 \;=\; 1 - \sum_{j \neq k} \left(\frac{\boldsymbol{v}_j^\mathsf{T} \boldsymbol{D}\, \boldsymbol{v}_k}{\lambda_k - \lambda_j} \right)^2.$$

We now focus on deriving an upper bound for the expected error of the first principal component, $\mathbb{E}\,\|\hat{\boldsymbol{v}}_1 - \boldsymbol{v}_1\|_2$, where the expectation is over noise. As a vector's ℓ^2 norm is always less than its ℓ^1 norm, we can bound the expectation of $\|\hat{\boldsymbol{v}}_1 - \boldsymbol{v}_1\|_1$ instead. By Proposition 1, we have to first order

$$\begin{aligned}
\mathbb{E}\,\|\hat{\boldsymbol{v}}_1 - \boldsymbol{v}_1\|_2 &\leq \mathbb{E}\,\|\hat{\boldsymbol{v}}_1 - \boldsymbol{v}_1\|_1 \\
&= \mathbb{E} \sum_{j \neq 1} \left| \frac{\boldsymbol{v}_j^\mathsf{T} \boldsymbol{D} \boldsymbol{v}_1}{\lambda_1 - \lambda_j} \right| \\
&\leq \mathbb{E} \left[\sum_{j \neq 1} \frac{1}{(\lambda_1 - \lambda_j)^2} \sum_{j \neq 1} (\boldsymbol{v}_j^\mathsf{T} \boldsymbol{D} \boldsymbol{v}_1)^2 \right]^{1/2}
\end{aligned} \tag{2.6}$$

where we have used the Cauchy-Schwarz Inequality to isolate the effects of the numerator and denominator. As the Pythagorean Theorem states that $\sum_j (\boldsymbol{v}_j^\mathsf{T} \boldsymbol{D} \boldsymbol{v}_1)^2 = \|\boldsymbol{D}\boldsymbol{v}_1\|^2$, we have, using the definition of matrix norm, that

$$\mathbb{E}\,\|\hat{\boldsymbol{v}}_1 - \boldsymbol{v}_1\|_2 \leq \mathbb{E}\,\|\boldsymbol{D}\| \left[\sum_{j \neq 1} \frac{1}{(\lambda_1 - \lambda_j)^2} \right]^{1/2}. \tag{2.7}$$

The norm of the covariance distortion $\|\boldsymbol{D}\|$ can be expanded from equation (2.4) as

$$\begin{aligned}
\|\boldsymbol{D}\| &= \|\hat{\boldsymbol{C}} - \boldsymbol{C}\| \\
&= (n-1)^{-1} \|(\boldsymbol{P} + \boldsymbol{E})(\boldsymbol{P} + \boldsymbol{E})^\mathsf{T} - \boldsymbol{P}\boldsymbol{P}^\mathsf{T}\| \\
&= (n-1)^{-1} \|\boldsymbol{P}\boldsymbol{E}^\mathsf{T} + \boldsymbol{E}\boldsymbol{P}^\mathsf{T} + \boldsymbol{E}\boldsymbol{E}^\mathsf{T}\| \\
&\leq (n-1)^{-1} \big(2\|\boldsymbol{P}\boldsymbol{E}^\mathsf{T}\| + \|\boldsymbol{E}\boldsymbol{E}^\mathsf{T}\| \big) \\
&\leq \frac{2}{n-1} \|\boldsymbol{P}\|\, \|\boldsymbol{E}\| + O(\|\boldsymbol{E}\|^2),
\end{aligned} \tag{2.8}$$

where the last inequality follows from the sub-multiplicativity of the matrix norm. Hence $\|\boldsymbol{D}\|$ is bounded by the product of $\|\boldsymbol{P}\|$ and $\|\boldsymbol{E}\|$ plus higher order error terms. Putting this result in equation (2.7), we have

Proposition 2. *With the notation established,*

$$\mathbb{E}\,\|\hat{\boldsymbol{v}}_1 - \boldsymbol{v}_1\|_2 \;\leq\; \mathbb{E}\, \frac{2}{n-1} \|\boldsymbol{P}\|\, \|\boldsymbol{E}\| \left[\sum_{j \neq 1} \frac{1}{(\lambda_1 - \lambda_j)^2} \right]^{1/2} + O(\|\boldsymbol{E}\|^2). \tag{2.9}$$

So far our analysis has been general, aside from dropping higher order terms in the perturbation expansion. We next analyze in turn each of the three terms on the right side of equation (2.9), $\|\boldsymbol{P}\|$, $\|\boldsymbol{E}\|$, and $\{\sum_{j\neq 1}(\lambda_1-\lambda_j)^{-2}\}^{1/2}$, and introduce assumptions where necessary to simplify. In particular, while $\|\boldsymbol{E}\|$ can be simplified in different ways depending on the assumptions made regarding noise, we will assume the noise model of the previous section to analyze this term.

The norm of $\boldsymbol{P}$ is easy to compute from the definition of the gene covariance matrix. As $\boldsymbol{C}$ is defined to equal $\boldsymbol{P}\boldsymbol{P}^{\mathsf{T}}/(n-1)$, we have that $\|\boldsymbol{P}\|^2=(n-1)\|\boldsymbol{C}\|$ from which follows

$$\|\boldsymbol{P}\|^2=(n-1)\lambda_1, \tag{2.10}$$

using the fact that $\boldsymbol{C}$ is positive semi-definite.

Next we turn to the norm of $\boldsymbol{E}$, which fundamentally represents the "strength" of the noise, or the "noise power," caused by sequencing at a shallow depth. Evaluating this quantity is more challenging, as from our noise model analysis of Section 2.1.2, each entry of $\boldsymbol{E}$ is a gaussian random variable with a different variance. Such matrices are studied in random matrix theory. For instance, corollary 4.2 of Tropp 2011 provides a "tail bound" for the probability that the norm of this random matrix exceeds a fixed quantity. In our notation, the tail bound states that

$$\Pr\{\|\boldsymbol{E}\|>\sqrt{t}\}\ <\ \min\{(n+g)\,\exp(-t/2\sigma^2),\ 1\},$$

where σ^2 is a "variance parameter" equal to

$$\sigma^2=\max\left\{\max_j\frac{1}{N}\sum_k P^0_{jk}(1-P^0_{jk}),\ \max_k\frac{1}{N}\sum_j P^0_{jk}(1-P^0_{jk})\right\}. \tag{2.11}$$

This tail bound is sufficient to bound the first two moments of $\|\boldsymbol{E}\|$ as shown in Section 4.3 of Tropp 2011. The second moment of $\|\boldsymbol{E}\|$ follows from the fact that the expectation of a non-negative random variable is one minus its cdf, so

$$\begin{aligned}
\mathbb{E}\,\|\boldsymbol{E}\|^2 &= \int_0^\infty \Pr\{\|\boldsymbol{E}\|>\sqrt{t}\}\,\mathrm{d}t\\
&\le \int_0^\infty \min\{(n+g)\,\exp(-t/2\sigma^2),\ 1\}\,\mathrm{d}t\\
&= 2\sigma^2\log(n+g)\ +\ \int_{2\sigma^2\log(n+g)}^\infty (n+g)\,\exp(-t/2\sigma^2)\,\mathrm{d}t\\
&= 2\sigma^2\log e(n+g).
\end{aligned}$$

This directly leads to a bound for expectation of $\|\boldsymbol{E}\|$. Since $\mathbb{V}\,\|\boldsymbol{E}\|=\mathbb{E}\,(\|\boldsymbol{E}\|^2)-(\mathbb{E}\,\|\boldsymbol{E}\|)^2$, we have that $\mathbb{E}\,\|\boldsymbol{E}\|\le(\mathbb{E}\,\|\boldsymbol{E}\|^2)^{1/2}$ from which

$$\mathbb{E}\,\|\boldsymbol{E}\|\le\sigma\{2\log e(n+g)\}^{1/2}. \tag{2.12}$$

The term in the square root depends on n and g but very weakly. For instance, taking the small values of $g=1000$ and $n=10$, the quantity $\{2\log e(n+g)\}^{1/2}$ is 3.97. On the other hand, for $g=n=10^5$, the term increases only to 5.14. Hence for values within an order of

magnitude of what we may encounter in practice, we incur little error by treating this term as constant.

We now simplify the variance parameter of equation (2.11). The variance parameter is the maximum of the largest row sum and the largest column sum of the variances in $\boldsymbol{E}$. As typically there are many more genes then samples, the largest column sum is greater than the largest row sum and therefore determines σ^2. More formally we have

Assumption 2. *As commonly $P_{ij} \ll 1$, assume that $P_{ij} \gg P_{ij}^2$. Additionally suppose that $\|\boldsymbol{P}\|_\infty < 1$. This latter assumption will be satisfied if n is small, say $n < 1/\sqrt{\lambda_1}$, as $\|\boldsymbol{P}\|_\infty \leq \sqrt{n}\|\boldsymbol{P}\|_2 \leq n\sqrt{\lambda_1}$.*

Then the variance parameter σ^2 reduces to

$$\sigma^2 \approx \max\left\{ \max_j \frac{1}{N}\sum_k P_{jk}^0,\ \max_k \frac{1}{N}\sum_j P_{jk}^0 \right\} = \max_k \frac{1}{N}\sum_j P_{jk}^0 = \frac{1}{N}. \tag{2.13}$$

Combining these results, we have shown

Proposition 3. *With Assumptions 1 and 2, the expectation of the norm of the error in the gene expression probabilities $\boldsymbol{E}$ satisfies*

$$\mathbb{E}\,\|\boldsymbol{E}\| \leq \frac{\kappa}{\sqrt{N}} \tag{2.14}$$

where κ is a constant ($\{2\log e(n+g)\}^{1/2}$) effectively independent of the dimensions of the matrix.

Finally, we analyze the third term on the right side of equation (2.9). Splitting the term into two sums yields

$$\begin{aligned}\left[\sum_{j\neq 1}\frac{1}{(\lambda_1-\lambda_j)^2}\right]^{1/2} &= \left[\left(\sum_{\substack{j\leq n\\ j\neq 1}}\frac{1}{(\lambda_1-\lambda_j)^2} + \sum_{j>n}\frac{1}{\lambda_1^2}\right)\right]^{1/2} \\ &\leq \left(\frac{n-1}{(\lambda_1-\lambda_2)^2} + \frac{g-n}{\lambda_1^2}\right)^{1/2},\end{aligned} \tag{2.15}$$

using the fact that $\lambda_1-\lambda_2 \leq \lambda_1-\lambda_j$ for all $j>1$. As discussed in the main text, typically the principal values "decay" rapidly so that the kth principal value spacing is large with respect to λ_k. This observation allows us to compare the relative magnitude of the two terms in equation (2.15) and motivates

Assumption 3. *Let $\delta\lambda_k \triangleq \min_j\{|\lambda_k-\lambda_j|\}$ be minimum distance between the kth principal value and any other principal value. Assume that $\boldsymbol{P}$ satisfies*

$$\sqrt{\frac{n}{g-n}} \ll \frac{\delta\lambda_k}{\lambda_k} \qquad \text{for all } k. \tag{2.16}$$

In general, we call matrices that satisfy this property well-separated.

In many cases of interest, $\delta\lambda_k = \lambda_k - \lambda_{k+1}$. Then equation (2.16) reduces to a simpler expression,

$$\frac{\lambda_{k+1}}{\lambda_k} \ll 1 - \sqrt{\frac{n}{g-n}} \qquad \text{for all } k.$$

Intuitively, this property states that each principal value is much smaller than the one that preceded it. This property can be checked in actual data in Figures S5C and S5D. With this assumption, we neglect the first term in the sum of equation (2.15) to obtain

$$\left[\sum_{j\neq 1}\frac{1}{(\lambda_j-\lambda_1)^2}\right]^{1/2} \leq \left(\frac{g-n}{\lambda_1^2}\right)^{1/2}. \tag{2.17}$$

We return to equation (2.9) and put all of these results together. If we apply the bounds of (2.10), (2.14), and (2.17) and drop the higher order error terms, we find

$$\mathbb{E}\,\|\hat{\boldsymbol{v}}_1-\boldsymbol{v}_1\|_2 \leq 2\kappa\left(\frac{g-n}{\lambda_1 N(n-1)}\right)^{1/2}.$$

Since Assumption 3 implies that $g \gg n$ and in practice $n \gg 1$, this inequality is approximately

$$\mathbb{E}\,\|\hat{\boldsymbol{v}}_1-\boldsymbol{v}_1\|_2 \leq \kappa\sqrt{\frac{1}{\lambda_1 N n}} \tag{2.18}$$

where we have absorbed constants into κ. This completes our derivation of Equation (2).

Remark It is natural to ask what happens when n is large, contrary to Assumption 2. In this case, the simplification used for the "variance parameter" in equation (2.13) no longer holds, and equation (2.14) is replaced with

$$\mathbb{E}\,\|\boldsymbol{E}\| \leq \frac{\kappa P_{\max}\sqrt{n}}{\sqrt{N}}$$

where $P_{\max}=\max_{ij}\{P_{ij}\}$. Consequently, the best bound our methods show for the error of the first principal component is

$$\mathbb{E}\,\|\hat{\boldsymbol{v}}_1-\boldsymbol{v}_1\|_2 \leq \kappa\sqrt{\frac{1}{\lambda_1 N}},$$

in place of equation (2.18).

2.1.4 Higher principal components

Similar reasoning shows that bound (2.18) can be adapted to apply to higher principal components. For principal component k, the previous analysis holds but equation (2.15) must be generalized to

$$\left[\sum_{j\neq k}\frac{1}{(\lambda_j-\lambda_k)^2}\right]^{1/2} \leq \left(\frac{n-1}{\min_j(\lambda_k-\lambda_j)^2} + \frac{g-n}{\lambda_k^2}\right)^{1/2}$$

With this modification, we find, using the same reasoning based on Assumption 3, that

$$\mathbb{E}\,\|\boldsymbol{v}_k-\boldsymbol{v}_k\|_2 \;\leq\; \kappa\,\frac{\sqrt{\lambda_1}}{\lambda_k}\sqrt{\frac{1}{nN}} \tag{2.19}$$

This is a conservative bound that will likely be sufficient for many applications. However, the bound can be improved as we show next.

The key idea to improving the bound is that $\boldsymbol{D}$ can be written as a sum of rank-one projections, many of which contribute only second order terms to the perturbation expansion of Proposition 1. We will remove these second order terms to find a "reduced perturbation" that we call $\boldsymbol{D}'$ (which has strictly smaller norm than $\boldsymbol{D}$) and use this perturbation to bound $\mathbb{E}\,\|\hat{\boldsymbol{v}}_k - \boldsymbol{v}_k\|$ through Proposition 1. As an illustration, consider the second principal component, $\boldsymbol{v}_2$, and its noisy counterpart, $\hat{\boldsymbol{v}}_2$. To first order,

$$\hat{\boldsymbol{v}}_2 - \boldsymbol{v}_2 = \sum_{j \neq 2} \frac{\boldsymbol{v}_j^\mathsf{T} \boldsymbol{D}\, \boldsymbol{v}_2}{\lambda_2 - \lambda_j} \boldsymbol{v}_j.$$

We expand the terms in $\boldsymbol{D} = \hat{\boldsymbol{C}} - \boldsymbol{C}$ via eigendecomposition to find

$$\hat{\boldsymbol{v}}_2 - \boldsymbol{v}_2 = \sum_{j \neq 2} \frac{\boldsymbol{v}_j^\mathsf{T} (\hat{\boldsymbol{C}} - \boldsymbol{C}) \boldsymbol{v}_2}{\lambda_2 - \lambda_j} \boldsymbol{v}_j = \sum_{j \neq 2} \left(\sum_{i=1}^{n} \hat{\lambda}_i \frac{\boldsymbol{v}_j^\mathsf{T} \hat{\boldsymbol{v}}_i \hat{\boldsymbol{v}}_i^\mathsf{T} \boldsymbol{v}_2}{\lambda_2 - \lambda_j} \boldsymbol{v}_j - \sum_{i=1}^{n} \lambda_i \frac{\boldsymbol{v}_j^\mathsf{T} \boldsymbol{v}_i \boldsymbol{v}_i^\mathsf{T} \boldsymbol{v}_2}{\lambda_2 - \lambda_j} \boldsymbol{v}_j \right).$$

In the inner sum, we incur an error of $O(\lambda_1 \|\boldsymbol{D}\|^2)$ if we choose to skip the $i = 1$ term. This is because both $\boldsymbol{v}_j^\mathsf{T} \hat{\boldsymbol{v}}_1$ and $\hat{\boldsymbol{v}}_1^\mathsf{T} \boldsymbol{v}_2$ are on the order of $\|\boldsymbol{D}\|$. For sufficiently large number of reads, this is smaller than the $i = 2$ term which is $O(\lambda_2 \|\boldsymbol{D}\|)$ as $\boldsymbol{v}_j^\mathsf{T} \hat{\boldsymbol{v}}_2$ is $O(\|\boldsymbol{D}\|)$ and $\boldsymbol{v}_2^\mathsf{T} \hat{\boldsymbol{v}}_2$ is $O(1 - \|\boldsymbol{D}\|^2)$. Discarding these second order terms, we have by this analysis

$$\hat{\boldsymbol{v}}_2 - \boldsymbol{v}_2 \approx \sum_{j \neq 2} \frac{\boldsymbol{v}_j^\mathsf{T} (\hat{\boldsymbol{C}}' - \boldsymbol{C}') \boldsymbol{v}_2}{\lambda_2 - \lambda_j} \boldsymbol{v}_j = \sum_{j \neq 2} \frac{\boldsymbol{v}_j^\mathsf{T} \boldsymbol{D}'\, \boldsymbol{v}_2}{\lambda_2 - \lambda_j} \boldsymbol{v}_j. \tag{2.20}$$

In this equation, $\boldsymbol{C}'$ and $\hat{\boldsymbol{C}}'$ are "reduced" matrices equal to $\sum_{i>1} \boldsymbol{v}_i \boldsymbol{v}_i^\mathsf{T}$ and $\sum_{i>1} \hat{\boldsymbol{v}}_i \hat{\boldsymbol{v}}_i^\mathsf{T}$ respectively, and $\boldsymbol{D}' \triangleq \hat{\boldsymbol{C}}' - \boldsymbol{C}'$ is the "reduced perturbation." In general, to bound the error of the kth principal component, the first $k-1$ rank one projections of $\boldsymbol{D}$ onto $\boldsymbol{v}_i \boldsymbol{v}_i^\mathsf{T}$ may be projected out with a loss of accuracy of only $O(\|\boldsymbol{D}\|^2)$. As the reduced perturbation $\boldsymbol{D}'$ has a smaller norm, we are able to obtain better bounds.

With equation (2.20), we continue like we did with the first principal component, replacing $\boldsymbol{D}$ with $\boldsymbol{D}'$ in our analysis. Equation (2.6) is modified to give

$$\begin{aligned}
\mathbb{E}\,\|\hat{\boldsymbol{v}}_2 - \boldsymbol{v}_2\|_2 &\leq \mathbb{E}\left[\sum_{j \neq 2} \frac{1}{(\lambda_j - \lambda_2)^2} \sum_{j \neq 2} (\boldsymbol{v}_j^\mathsf{T} \boldsymbol{D}' \boldsymbol{v}_2)^2 \right]^{1/2} \\
&\leq \mathbb{E}\,\|\boldsymbol{D}'\| \left[\sum_{j \neq 2} \frac{1}{(\lambda_j - \lambda_2)^2} \right]^{1/2} \\
&\leq \mathbb{E}\left[\frac{2}{n-1} \|\boldsymbol{P}'\|\, \|\boldsymbol{E}\| + O(\|\boldsymbol{E}\|^2) \right] \left[\sum_{j \neq 2} \frac{1}{(\lambda_j - \lambda_2)^2} \right]^{1/2}
\end{aligned} \tag{2.21}$$

Here $\boldsymbol{P}'$ is the "reduced" data, found by forming the singular value decomposition of $\boldsymbol{P}$ without the first component, *i.e.* $\sum_{i>1} \sqrt{\lambda_i}\, \boldsymbol{v}_i \boldsymbol{w}_i^\mathsf{T}$, and row mean centering the result. By construction, $\boldsymbol{P}'\boldsymbol{P}'^\mathsf{T}$ and $\boldsymbol{P}\boldsymbol{P}^\mathsf{T}$ share the same eigenvectors.

Now consider the two terms in equation (2.21). The first term in brackets depends on $\|\boldsymbol{P}'\|$ and $\mathbb{E}\,\|\boldsymbol{E}\|$. Equation (2.14) describing $\mathbb{E}\,\|\boldsymbol{E}\|$ is unchanged but equation (2.10) is now replaced with

$$\|\boldsymbol{P}'\|^2 = (n-1)\lambda_2. \tag{2.22}$$

The second term in brackets is

$$\begin{aligned}\left[\sum_{j\neq 2}\frac{1}{(\lambda_j-\lambda_2)^2}\right]^{1/2} &= \left[\left(\sum_{\substack{1\leq j\leq n\\ j\neq 2}}\frac{1}{(\lambda_j-\lambda_2)^2} + \sum_{j>n}\frac{1}{\lambda_2^2}\right)\right]^{1/2}\\ &\leq \left(\frac{n-1}{\min_j(\lambda_2-\lambda_j)^2} + \frac{g-n}{\lambda_2^2}\right)^{1/2}\end{aligned}$$

Now using the assumption that $\boldsymbol{P}$ is well-separated, this inequality is approximately

$$\left[\sum_{j>2}\frac{1}{(\lambda_j-\lambda_2)^2}\right]^{1/2} \leq \left(\frac{g-n}{\lambda_2^2}\right)^{1/2}. \tag{2.23}$$

Substituting equation (2.22) and equation (2.23) into equation (2.21) and dropping higher order terms, we have

$$\mathbb{E}\,\|\hat{\boldsymbol{v}}_2-\boldsymbol{v}_2\|_2 \leq \frac{\kappa}{\sqrt{nN\lambda_2}}.$$

This technique can be applied iteratively. For the kth principal component, the first $k-1$ rank one projections of $\boldsymbol{D}$ onto $\boldsymbol{v}_i\boldsymbol{v}_i^{\mathsf{T}}$ can be neglected, so that the norm of the "reduced" data $\boldsymbol{P}'$ is $\{(n-1)\lambda_k\}^{1/2}$. Hence the error in the kth principal component is

$$\mathbb{E}\,\|\hat{\boldsymbol{v}}_k-\boldsymbol{v}_k\|_2 \leq \frac{\kappa}{\sqrt{nN\lambda_k}}. \tag{2.24}$$

2.2 Gene expression modules enhance principal value separation in a simple model

In the main text we show that principal value separation determines the accuracy of transcriptional program extraction at low read-depths. Through a broad survey of gene expression datasets, we find that favorable principal value separation is common in biological data allowing mRNA-seq at a drastically increased scale. In this section, we study a simple gene expression model to ask how modularity, a core structural property of biological systems, might impart principal value separation and noise tolerance to gene expression data. The relationship between gene expression modules and principal value separation is of fundamental interest because transcriptional regulatory networks are commonly organized into regulatory modules, groups of covarying genes. In fact the covariance matrices of both the Shen et al. and Zeisel et al. datasets contain coherent gene expression blocks that suggest an underlying modular architecture (Figure 4B and S4). Within the context of a simple model, we rigorously show that principal value separation is enhanced by such a modular architecture; principal value separation scales directly with module size and the magnitude of gene expression covariance within modules. As such, gene expression modules might endow biological systems with an inherent tolerance to shallow profiling.

Gene expression model We consider a gene expression covariance matrix, $\boldsymbol{C}$, that has the block diagonal structure shown in Figure S3B. The matrix contains two blocks of size $b \times b$, and each block represents a module of covarying genes. Genes within each red block have a positive covariance, and the two blocks have relative negative covariance represented by green. For mathematical convenience, we consider all blocks to share a constant within-block gene expression covariance q. The gene expression variances m_i within each block are assumed to be constant, but differ between the blocks to avoid module degeneracy. The between-block covariances are likewise constant and equal to r. We note that for a gene expression model with two mutually exclusive gene expression modules (*i.e.* when module 1 is "on" module 2 is "off"), $r \leq 0$ because covariance is calculated following mean centering of the raw gene expression data. We also assume that $q \geq 0$, $m_1 > m_2$, $m_1 > q$, $m_2 > q$. Finally, we assume that $q > |r|$ to ensure that the two blocks of genes are distinct.

Solution to the model In this simple model, we now determine the factors that influence the separation of the principal values of the underlying data. The principal values, denoted by λ_i, are defined as the eigenvalues of $\boldsymbol{C}$ and for this model can be calculated analytically. There are $2b$ eigenvalues in total: $b-1$ eigenvalues equal to $m_1 - q$, another $b-1$ eigenvalues equal to $m_2 - q$, and two eigenvalues given by

$$\{\lambda_1, \lambda_2\} \;=\; \overline{m} + (b-1)q \pm \sqrt{(br)^2 + \{\delta m\}^2}, \tag{2.25}$$

with the notation $\overline{m} = (m_1 + m_2)/2$ and $\delta m = (m_1 - m_2)/2$. When q and r both equal zero, all gene-gene covariances are zero and the system has eigenvalues m_1 and m_2, both with multiplicity b. For nonzero q and r, the non-degenerate eigenvalues separate from the degenerate eigenvalues. In this case, λ_1 is the largest eigenvalue of $\boldsymbol{C}$ and, for sufficiently large q, λ_2 is the second largest eigenvalue of $\boldsymbol{C}$ (Figure S3C). We note that the noise tolerance of λ_1 is of special interest because its associated eigenvector, the first principal component of the gene expression data, identifies the gene expression modules in the system (Figure S3B). The entries of this eigenvector, $\mathbf{pc}_1$, are positive for genes within one module and negative for genes within the other module.

As described in the main text, the noise tolerance of $\mathbf{pc}_1$ depends upon the spacing between λ_1 and all other eigenvalues of $\boldsymbol{C}$. These can be calculated directly as

$$\lambda_1 - \lambda_2 = 2\sqrt{(br)^2 + \{\delta m\}^2} \tag{2.26a}$$

$$\lambda_1 - (m_1 - q) = \sqrt{(br)^2 + \{\delta m\}^2} - \delta m + bq \tag{2.26b}$$

$$\lambda_1 - (m_2 - q) = \sqrt{(br)^2 + \{\delta m\}^2} + \delta m + bq. \tag{2.26c}$$

These quantities are depicted as a function of q in Figure S3C for a two-module, four gene system.

The features that improve the noise tolerance of principal component recovery Examination of the the principal value separations leads to two conclusions. First, increasing the within-module covariance q increases the separation between λ_1 and λ_i for $i > 2$ (equations 2.26b, 2.26c). Secondly, this effect scales with the size of the gene expression modules. While both the covariance term q and the variance term m_i contribute to principal value separation, only the impact of the covariance term scales with block size b (as bq). Hence for large modules (*i.e.* large b), the covariance terms may contribute significantly to

principal value separation. We conclude that gene expression modularity directly increases the separation between λ_1 and all other principal values, and thus enhances the ability to extract principal component 1 at low read depth. (We note that our model with covariance q fixed for all blocks can be generalized to allow for differing covariance parameters within blocks and still yields qualitatively similar results.)

Finally, the impact of module size on principal value separation can be seen directly in a generalized model where more than two gene expression modules are allowed. In Figure S3A, we analyze a series of covariance matrices where the number of genes is held constant, but the number of gene expression modules is increased. In these calculations, the covariance terms q and r are held constant ($q = 40, r = -8$) and m_i span a constant range, $80 < m_i < 200$. For constant q and r, the spacing between the largest eigenvalue and all others eigenvalues ($\lambda_1 - \lambda_i$) increases as module size increases. A system with two modules has significantly increased principal value separation when compared with even a four module system. Due to this scaling, large gene expression modules might significantly enhance principal value separation and therefore noise tolerance in biological data.

3 Supplemental Experimental Procedures

Alignment of sequencing reads and quantification of read counts for public mRNA-seq datasets

Raw mRNA-seq reads were obtained from the Gene Expression Omnibus. mRNA-seq datasets used in Figure S5A are from Shen et al. 2012; Treutlein et al. 2014; Shalek et al. 2013; Kumar et al. 2014; Chen et al. 2012; Pollen et al. 2014. The reads from these studies were aligned to either human *hg19* or mouse *mm9* exomes. Exome files were constructed based upon the transcriptome annotations and gene feature files (gff) available from the UCSC genome browser (Kent et al. 2002). Open reading frames encoding rDNA, transposable elements, or other non-protein coding features were not included in the exome. Reads were aligned to exomes using Bowtie2 v 2.1.0 (Langmead and Salzberg 2012) using the following options -D 25 -R 3 -N 1 -L 20 -i S 1 0.50 local.

Following alignment, the data was preprocessed prior to analysis. This was accomplished by normalizing raw per gene read counts by the total number of reads collected for a given sample, ensuring that the normalized reads of one experiment sum to one.

For the analysis of Zeisel et al., we used the transcript counts reported on the Linnarsson Lab website.

Simulated shallow sequencing through down-sampling of reads

A computational downsampling procedure was applied to simulate the impact of reduced read depth on public mRNA-seq datasets (the Zeisel et al. data required a different method; see the next experimental procedure). Read counts for the deep mRNA-seq experiments were normalized by dividing the number of reads mapped to a gene by the total number of mapped reads in that experiment, generating a multinomial probability mass function. For a given simulated read depth, we model the sequencing process by drawing N reads, with replacement, from this multinomial distribution.

We sample with replacement because the number of molecules within an mRNA-seq library, $\sim 10^{12}$ (McIntyre et al. 2011), is much larger than the number of reads being sequenced, $\sim 10^{7}$ (Shen et al. 2012), effectively making each sequencing event independent of others. To accelerate the computation at simulated depths over one million reads, read counts were estimated directly with a Poisson distribution. Similar downsampling procedures are frequently use to model read depth reductions and associated measurement noise (Robinson and Storey 2014; Pollen et al. 2014).

Simulated shallow sequencing through down-sampling of transcripts for Zeisel et al. dataset

As the Zeisel et al. data contains unique molecular identifiers (UMIs) which allow for the direct quantification of transcripts, the previously described downsampling procedure was modified for this dataset. First, 15,000 transcripts were sampled with replacement for each cell (as previously described) to obtain gene expression profiles with constant transcript coverage per sample. We sampled 15,000 transcripts as that is approximately the average number of unique transcripts observed per cell in Zeisel et al. To simulate low coverage data, we sampled a desired number of reads from this reference distribution without replacement.

Saturating expression levels of outlying genes

Following downsampling, the largest 1% of all gene expression values (based on read counts) were set to the value of the 99^{th} percentile of the data. This saturation was performed to diminish the impact of extreme outliers on subsequent data analysis as PCA is known to be sensitive to such outliers. Following saturation, data was renormalized to preserve the equal weighting of each experiment. We found in practice that outlier filtering was important for preserving biological structure and in fact was required for biological replicates to cluster together in the mouse tissue dataset. The saturation threshold for Kumar et al. 2014 was an exception to the 1% threshold, it was set to 2.25% to ensure biological replicates clustered together. Read counts were used as the fundamental gene expression unit in the analysis for simplicity in theoretical modeling and convenience during the simulated down-sampling procedure. Similar results were obtained in FPKM units where read counts are normalized for gene length.

For the Zeisel et al. dataset, after downsampling and before principal components analysis, we removed the top 15 varying genes. We found that this was necessary for recapitulating the original study's classification of cell types at full transcript coverage.

Evaluation of Equation 1

Evaluation of Equation 1 requires the deep principal components, deep principal values, and the deep and shallow data covariance matrices. The deep principal components and principal values were determined for each dataset directly from the deep normalized read count data. $\hat{\mathbf{C}}$ was then calculated on read count data generated through the simulated down-sampling procedure described above. At each read depth, Equation 1 was evaluated on twenty separate instances of $\hat{\mathbf{C}}$, and the mean principal component error was reported as a percent of the theoretical maximum error ($\sqrt{2}$).

Projecting gene expression profiles onto the principal components

Classification plots show the principal component coefficients for each gene expression profile (from either a bulk mRNA-seq sample or single cell). These coefficients represent the amount of variance along the axis defined by the respective principal component, or the projection of the expression profile onto a principal component. These coefficients are computed by taking the dot product of the gene expression profile and the principal component. When simulating low coverage mRNA-seq, the noisy, simulated gene expression profile is projected onto the principal components computed from the noisy, simulated gene expression data.

Zeisel et al. cell type classification accuracy

For classification of single cells from Zeisel et al. at a simulated depth, each sampled transcriptional profile was compared to three reference transcriptional profiles. The reference transcriptional profiles were computed by averaging the full depth transcriptional profile of each cell type as classified by Zeisel et al. Each downsampled cell was then assigned the cell type label of the most similar reference profile. False positives correspond to mismatches between the assigned cell type and the cell type from Zeisel et al. at full depth.

Cell type classification by nonlinear dimensionality reduction

Figures S2K and S2L were generated by downsampling data from Kumar et al. and Shalek et al. as described above, followed by dimensionality reduction with t-SNE and LLE. t-SNE was applied through the scikit-learn Python package version 2.7.6 (Pedregosa et al. 2011) and LLE was implemented following Roweis and Saul 2000.

Simple gene expression model

Simulated covariance matrices were generated for a system with six gene expression modules. Module size was drawn from an exponential distribution and modules sorted for increasing size. Within-module covariance was set to a uniform value (q, ranging from 10 to 80) and between-module covariance (r) was set to a constant for simplicity (-10).

Analysis and simulated downsampling of microarray data

Microarray data were downloaded from Gene Expression Omnibus (Edgar, Domrachev, and Lash 2002). To minimize the effect of platform variability, one type of microarray platform was selected for each species. We chose Affymetrix Yeast Genome S98 Array, Affymetrix Mouse Genome 430 2.0 Array, and Affymetrix Human Genome U133 Plus 2.0 Array because they had the largest number of datasets containing at least 20 samples. Log-transformed datasets were removed to ensure that each dataset was preprocessed in the same way. After filtering, 20 datasets for *Saccharomyces cerevisiae*, 106 datasets for *Mus musculus* and 226 for *Homo sapiens* remained. Each dataset was normalized so that the gene expression values of each sample sum to one. Further, as with the mRNA-seq datasets, we saturate expression levels at the 99% percentile to handle extreme expression outliers. To generate simulated mRNA-seq data from microarray experiments, we used the normalized and filtered gene expression matrix as input into our down-sampling procedure previously

described. The normalized gene expression matrix was used as a multinomial probability distribution and this distribution was sampled to generate simulated mRNA-seq data at different read depths.

Equation 2: fitting the constant κ and cross validation

To fit κ in Equation 2, we first partitioned the microarray data into a training set and a cross validation set, each consisting of half the datasets. We simulated shallow sequencing on the microarray data within the training set at ten values between 10^3 and 10^7 reads to obtain PCA error for each dataset. The constant κ from the Equation 2 was determined by fitting (with a linear regression) the simulated PCA error to the analytical prediction of Equation 2.

To demonstrate that the relationship is predictive, we simulated shallow sequencing on the remaining 50% of the microarray datasets at all ten depths and compared the predicted values with those observed from simulation. We further used the four mRNA-seq datasets for mouse and human (which were not used for fitting) as additional cross-validation. For each species, only one value of κ is calculated globally, and this value is used for all principal components, read depths, and datasets (for humans, $\kappa = 71.25$ and for mice, $\kappa = 69.30$).

Principal values and error in non-negative matrix factorization

Non-negative matrix factorization (NMF) was performed on normalized read count matrices generated from microarray database. Three "deep" NMF vectors were computed from the original data and three "shallow" NMF vectors were computed from simulated mRNA-seq data with 45,000 reads. The pair of computations shared the same random initialized state. Each shallow NMF vector from the simulated shallow mRNA-seq data was matched with a corresponding deep vector. Our algorithm determined matches by finding the one-to-one mapping that minimized the summed squared differences between the deep and corresponding shallow NMF vectors. The normalized error was computed as the magnitude of the difference between the matched NMF parts divided by the magnitude of the deep NMF part. The sum of the normalized errors was computed for three different initialization states. The median of the summed errors was used in Figure S5B.

Supplemental References

Anders, S. and Huber, W. (2010). "Differential expression analysis for sequence count data". *Genome Biology* 11.10.

Brennecke, P. et al. (2013). "Accounting for technical noise in single-cell RNA-seq experiments". *Nature Methods* 10.11, pp. 1093–1095.

Chen, R. et al. (2012). "Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes". *Cell* 148.6, pp. 1293–1307.

Daley, T. and Smith, A. D. (2014). "Modeling genome coverage in single-cell sequencing". *Bioinformatics* 30.22, pp. 3159–3165.

Ding, B. et al. (2015). "Normalization and noise reduction for single cell RNA-seq experiments". *Bioinformatics* 31 (13), pp. 2225–7.

Edgar, R., Domrachev, M., and Lash, A. E. (2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". *Nucleic Acids Research* 30.1, pp. 207–210.

Grün, D., Kester, L., and van Oudenaarden, A. (2014). "Validation of noise models for single-cell transcriptomics". *Nature Methods* 11.6, pp. 637–640.

Islam, S. et al. (2014). "Quantitative single-cell RNA-seq with unique molecular identifiers". *Nature Methods* 11.2, pp. 163–166.
Kent, W. J. et al. (2002). "The Human Genome Browser at UCSC". *Genome Research* 12.6, pp. 996–1006.
Kumar, R. M. et al. (2014). "Deconstructing transcriptional heterogeneity in pluripotent stem cells". *Nature* 516.7529, pp. 56–61.
Langmead, B. and Salzberg, S. L. (2012). "Fast gapped-read alignment with Bowtie 2". *Nature Methods* 9.4, pp. 357–359.
Liu, Y., Zhou, J., and White, K. P. (2014). "RNA-seq differential expression studies: more sequence or more replication?" *Bioinformatics* 30.3, pp. 301–304.
Marioni, J. C. et al. (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays". *Genome research* 18.9, pp. 1509–1517.
McIntyre, L. M. et al. (2011). "RNA-seq: technical variability and sampling". *BMC Genomics* 12.1, p. 293.
Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research* 12, pp. 2825–2830.
Pollen, A. A. et al. (2014). "Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex". *Nature Biotechnology* 32.10, pp. 1053–1058.
Robinson, D. G. and Storey, J. D. (2014). "subSeq: determining appropriate sequencing depth through efficient read subsampling". *Bioinformatics* 30.23, pp. 3424–3426.
Roweis, S. T. and Saul, L. K. (2000). "Nonlinear Dimensionality Reduction by Locally Linear Embedding". *Science* 290.5500, pp. 2323–2326.
Shalek, A. K. et al. (2013). "Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells". *Nature* 498.7453, pp. 236–40.
Shankar, R. (2012). *Principles of Quantum Mechanics*. Springer Science & Business Media.
Shen, Y. et al. (2012). "A map of the cis-regulatory sequences in the mouse genome". *Nature* 488.7409, pp. 116–120.
Shiroguchi, K. et al. (2012). "Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes". *Proceedings of the National Academy of Sciences* 109.4, pp. 1347–1352.
Stewart, G. W. and Sun, J. (1990). *Matrix Perturbation Theory*. Academic Press.
Tarazona, S. et al. (2011). "Differential expression in RNA-seq: A matter of depth". *Genome Research* 21.12, pp. 2213–2223.
Treutlein, B. et al. (2014). "Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq". *Nature* 509.7500.
Tropp, J. A. (2011). "User-Friendly Tail Bounds for Sums of Random Matrices". *Foundations of Computational Mathematics* 12.4, pp. 389–434.
Vallejos, C. A., Marioni, J. C., and Richardson, S. (2015). "BASiCS: Bayesian Analysis of Single-Cell Sequencing Data". *PLoS Comput Biol* 11.6, e1004333.