F. Theory Supplement

Here we derive Eqs. (1)-(3) in Fig. 2G of the main text, showing how the main sources of technical noise encountered in inDrop RNA sequencing and other transcriptomic methods affect gene expression variability and correlations observed in single cell sequencing data. We develop a statistical test based on the results presented here to identify highly variable genes.

F.1 A model for technical noise

Let n_i be the number of mRNA molecules in a given cell that correspond to a particular gene i, with a distribution $P_{\text{bio}}(n_i)$ across all cells being analyzed. Also, let m_i be the number of UMI-filtered mapped (UMIFM) sequencing reads mapping to the same gene i, with distribution $P_{\text{obs}}(m_i)$ obtained from the sequencing run. Ultimately we want to use the distribution of UMIFM reads $P_{\text{obs}}(m_i)$ to infer properties of $P_{\text{bio}}(n_i)$ such as its average $E[n_i]$, its variance Var(n_i), its coefficient of variation $\text{CV}_i = \sqrt{\text{Var}(n_i)}/E[n_i]$, and its modality. We also want to infer properties of the joint distributions of multiple genes – for example, the strength of the correlation between gene i and gene j, which we can calculate from the pairwise distributions $P_{\text{bio}}(n_i, n_j)$. The challenge in developing a model of noise in single cell transcriptomics is to explain how the joint distribution of the UMIFM read counts of all genes $P_{\text{obs}}(\{m_i\})$ relates the joint distribution of the transcript counts of these genes $P_{\text{bio}}(\{n_i\})$. From $P_{\text{obs}}(\{m_i\})$ we can extract any marginal distribution of interest.

To relate the set of read counts across all genes $\{m_i\}$ to the set of transcript counts $\{n_i\}$, we start with the chain rule,

$$P_{\text{obs}}(\{m_i\}) = \sum_n P_{\text{bio}}(\{n_i\}) \prod_i Q_i(m_i|n_i), \tag{S1}$$

where $\{Q_i(m_i|n_i)\}$ are the conditional probability distributions for observing $\{m_i\}$ UMIFM reads given $\{n_i\}$ transcripts. The extent to which $P_{\text{obs}}(\{m_i\})$ reflects the biology depends on the structure of $\{Q_i(m_i|n_i)\}$. Note that, implicit in our notation for Q_i, we have assumed that transcripts are sampled independently of one another within each droplet, although there may be variation between droplets. This assumption is supported by the excellent fit of the sensitivity curve (Fig. 2E) assuming independent and random sampling of mRNA transcripts (see Supplementary Methods section on Sensitivity). Thus the number of UMIFM reads m_i for a given gene should depend only on the actual number of transcripts n_i for that gene and not on the number of UMIFM reads m_j or transcripts n_j for any other gene.

To construct $Q_i(m_i|n_i)$, we consider the type of noise apparent in our system. Previous studies (Grun et al., 2014; Brennecke et al., 2013; Islam et al., 2014) assumed that sequenced transcripts are sampled from the pool of all transcripts in a cell according to Poisson statistics, i.e. through random sampling with replacement. These studies were motivated by the observation that the majority of genes in a sample follow a Poisson noise relationship, $\text{CV}^2 \sim 1/(\text{mean})$, with a baseline additive technical noise. Here, we make a subtle correction to these studies by noting that sampling of mRNA transcripts occurs *without* replacement, giving rise to a Binomial, not Poisson, distribution. A Binomial sampling is expected when sequencing depth is sufficient such that the number of reads per cell is limited only by the capture efficiency of the transcripts; if however the number of sequenced reads is too low to capture the full complexity of the RNA-Seq library (for example, if the number of cells sequenced in a single run is very high), the limited number of reads per cell gives rise to a Hypergeometric distribution instead of a Binomial, but with similar results. With this correction we are able to show from first principles how a baseline noise arises, and that it not only additively complements the gene CVs but also multiplicatively amplifies existing biological variation.

The choice of a Binomial distribution is motivated by the observation that the number of UMIFM reads m_i for a given gene from a given cell can only under-estimate the true number of transcripts n_i. Thus m_i is drawn from a Binomial distribution characterized by a sampling efficiency β corresponding to the probability of any individual mRNA molecule being sampled. Since β can fluctuate between droplets independently of m_i, Q_i is the Binomial distribution marginalized over fluctuations in β, viz.

$$Q_i(m_i|n_i) = \int d\beta \; \xi(\beta) Bi(m_i; n_i, \beta), \tag{S2}$$

where $Bi(m_i; n_i, \beta)$ is the Binomial distribution,

$$Bi(m_i; n_i, \beta) = \binom{n_i}{m_i} \beta^{m_i} (1-\beta)^{n_i - m_i},$$

and $\xi(\beta)$ is the distribution across droplets of sampling efficiencies β. Note that Eqs. (S1) and (S2) assume that all transcripts within the same droplet are sampled with the same efficiency β, which may not be true (for example) if some transcripts within the same cell are more or less accessible to primer capture than others. These equations also ignore other sources of noise such as ambiguities in mapping UMIFM reads to genes and rare events in which two cells are present in the same droplet. Despite these limitations, from Eqs. (S1) and (S2) one may derive predictions for the observed gene CVs and correlations that agree well with trends seen in the data, and which provide an intuitive explanation for sources of variation in the data.

Having laid out the basic structure and assumptions of our noise model, we now use them to relate variability and correlations in the number of UMIFM reads for genes to those properties of the actual number of transcripts. We also explore how normalizing the data affects our results.

F.2 Key results

From the noise model in section F.1, we find that technical noise amplifies existing biological variation of a gene's abundance across cells and weakens correlations between genes. Here we formalize these intuitive behaviors through equations that relate the biological CV and pairwise correlation strength with their experimentally observable counterparts.

F.2.1 Technical noise amplifies above-Poisson biological variation

Equations (S3) and (S4) below present the key relationships describing the observed CV of gene expression. The first equation holds for unnormalized data; the second equation refers to data normalized by the total counts per cell (as defined below), with the normalized UMIFM counts denoted as $\hat{m}$. The normalization procedure reduces technical noise in the efficiency β, but it undesirably inflates the CV estimates for each gene by the cell-to-cell variability in total mRNA content $N = \sum_i n_i$, which may reflect fluctuations in cell size or cell cycle stage. Eq. (S4) is accurate for genes whose transcript abundances are independent of the total number of transcripts in a cell, N, an assumption that is almost certainly incorrect for genes that correlate strongly with the cell cycle. In this section we drop the subscript i from all equations since they apply to genes individually rather than jointly.

No normalization:

$$\mathrm{CV}_m^2 - \frac{1}{E[m]} = \left(\mathrm{CV}_n^2 - \frac{1}{E[n]}\right)\left(1 + \mathrm{CV}_\beta^2\right) + \mathrm{CV}_\beta^2 \tag{S3}$$

Total count normalization:

$$\mathrm{CV}_{\hat{m}}^2 - (1 + \mathrm{CV}_M^2)(1 + \mathrm{CV}_{1/N}^2)\frac{1}{E[\hat{m}]} = \left(\mathrm{CV}_n^2 - \frac{1}{E[n]}\right)\left(1 + \mathrm{CV}_{1/N}^2\right) + \mathrm{CV}_{1/N}^2 \tag{S4}$$

Technical noise is represented in both Eqs. (S3) and (S4) by variability CV_β in the sampling efficiency of the method. Eq. (S4) includes, as we would expect, variability CV_M across cells or control droplets in the total number of UMIFM counts, $M = \sum_i m_i$. Note that M and CV_M are empirical quantities that can be calculated directly from the data. Eq. (S4) also captures variability in the total number N of mRNA transcripts originally present in those cells, in the form of $\mathrm{CV}_{1/N}$.

We begin the derivation of Eqs. (S3) and (S4) the same way. Both equations follow from the Laws of Total Expectation and Total Variance applied to the conditional means and variances of (normalized) read counts m ($\hat{m}$), conditioned on the actual number of transcripts n and the sampling efficiency β. For Binomial sampling, these conditional moments are as follows:

$$E[m|n,\beta] = \beta n$$
$$\mathsf{Var}(m|n,\beta) = \beta(1-\beta)n.$$

We now calculate the unconditional moments $E[m]$ and $\mathsf{Var}(m)$ in terms of these conditional ones using the Laws of Total Expectation and Total Variance:

$$E[m] = E_{n,\beta}[E[m|n,\beta]], \tag{S5}$$

$$\mathsf{Var}(m) = E_{n,\beta}[\mathsf{Var}(m|n,\beta)] + \mathsf{Var}_{n,\beta}(E[m|n,\beta]), \tag{S6}$$

where $E_{n,\beta}[g(n,\beta)] = E_n[E_\beta[g(n,\beta)]]$ is the expected value of a function $g(n,\beta)$ over the distributions of n and β, and $Var_{n,\beta}(g) = E_{n,\beta}[g^2(n,\beta)] - E^2_{n,\beta}[g(n,\beta)]$. We obtain:

$$E[m] = E[\beta]E[n]$$
$$\mathsf{Var}(m) = E[\beta]E[n] - E[\beta^2]E[n] + E[\beta^2]E[n^2] - E[\beta]^2E[n]^2$$

We arrive at Eq. (S3) by evaluating $\mathrm{CV}_m^2 = \text{Var}(m)/E[m]^2$ and simplifying the result using the identity $\dfrac{E[\beta^2]}{E[\beta]^2} = 1 + \mathrm{CV}_\beta^2$.

Next we turn to total count normalization [Eq. (S4)]. In total count normalization, normalized read counts $\hat{m}$ are calculated in each droplet as follows:

$$\hat{m} \equiv m\frac{E[M]}{M}$$

where $M = \sum_i m_i$ is the total number of UMIFM reads (i.e., counts or library size) for a given cell, and $E[M]$ is the average of those totals across all cells. Because each transcript count m_i is binomially distributed conditional on β and n_i, M is binomially distributed conditional on β and $N = \sum_i n_i$. Since N is large, we say that M in each individual droplet is approximately its conditional expectation $E[M|\beta, N] = \beta N$. With this approximation, and taking β and N to be independent,

$$\hat{m} = m\frac{E[\beta]}{\beta}\frac{E[N]}{N} = m\frac{E[\beta]}{\beta}R \tag{S7}$$

To simplify subsequent algebra we define the random variable $R \equiv E[N]/N$. Proceeding as before, the conditional moments for $\hat{m}$ are

$$\begin{aligned} E[\hat{m}|n,\beta,R] &= \left(\frac{E[\beta]}{\beta}R\right)E[m|n,\beta] = E[\beta]Rn \\ \text{Var}[\hat{m}|n,\beta,R] &= \left(\frac{E[\beta]^2}{\beta^2}R^2\right)\text{Var}(m|n,\beta) = E[\beta]^2(\beta^{-1}-1)R^2 n. \end{aligned}$$

The quantity $\text{Var}[\hat{m}|n,\beta,R]$ depends on β^{-1}. Using Equation (S6) we conclude that the unconditional variance will depend on the inverse moment $E\left[\beta^{-1}\right]$, which we can approximate using a power series expansion,

$$\begin{aligned} E\left[\beta^{-1}\right] &= \frac{1}{E[\beta]}E\left[\frac{1}{1+(\beta-E[\beta])/E[\beta]}\right] \\ &= \frac{1}{E[\beta]}E\left[\sum_{k=0}^{\infty}(-1)^k\frac{(\beta-E[\beta])^k}{E[\beta]}\right] \\ &= \frac{1}{E[\beta]}\left(1+\mathrm{CV}_\beta^2\right) + O\left(\frac{E[(\beta-E[\beta])^3]}{E[\beta]^3}\right). \end{aligned}$$

In the final line above, we note that the quadratic term in the power series is CV_β^2. The higher order terms depend on the third and higher mean-normalized central moments of β, which we can safely ignore if the noise in β is small. Armed with this approximation for $E\left[\beta^{-1}\right]$, we find that

$$E[\hat{m}] = E[\beta]E[R]E[n] \tag{S8}$$

$$\begin{aligned} \text{Var}(\hat{m}) \approx\ & E[\beta]E[R^2]E[n]\big(1+\mathrm{CV}_\beta^2\big) - E[\beta]^2E[R^2]E[n] \\ & + E[\beta]^2E[R^2]E[n^2] - E[\beta]^2E[R]^2E[n]^2. \end{aligned} \tag{S9}$$

Now dividing Eq. (S9) by the square of Eq. (S8) gives

$$\mathrm{CV}_{\hat{m}}^2 - \frac{E[R](1+\mathrm{CV}_R^2)(1+\mathrm{CV}_\beta^2)}{E[\hat{m}]} = \left(\mathrm{CV}_n^2 - \frac{1}{E[n]}\right)\left(1+\mathrm{CV}_R^2\right) + \mathrm{CV}_R^2$$

To recover Eq. (S4), we make use of the following consequences of the equalities $R = E[N]/N$ and $M = \beta N$. First, we note that $\mathrm{CV}_R^2 = \mathrm{CV}_{1/N}^2$. Second, we can repurpose our power series expansion above for N instead of β and see that $E[R] = E[N]E[N^{-1}] \approx 1 + \mathrm{CV}_N^2$. Finally, since β and N are independent, we can say that $(1 + \mathrm{CV}_N^2)(1 + \mathrm{CV}_\beta^2) = (1 + \mathrm{CV}_{\beta N}^2) = 1 + \mathrm{CV}_M^2$.

F.2.2 Technical noise weakens observed gene-gene correlations

Technical noise may either weaken pairwise correlations between genes, or spuriously generate correlations through normalization. If two genes are sampled unevenly, their relationship in the sample may look quite different from their relationship in the original pool. Moreover, correlation is sensitive to scale – two low-abundance genes are much more likely to seem uncorrelated than two highly abundant genes. The equation we develop here helps us understand more formally how sampling and noise in sampling weaken correlations that we observe between genes through their UMIFM read counts $\mathrm{corr}(\hat{m}_i, \hat{m}_j)$. Here we consider only the case of total count normalization. We begin with the definition of the correlation coefficient,

$$\mathrm{corr}(\hat{m}_i, \hat{m}_j) = \frac{\mathrm{Cov}(\hat{m}_i, \hat{m}_j)}{\sqrt{\mathsf{Var}(\hat{m}_i)\mathsf{Var}(\hat{m}_j)}},$$

and rewrite this expression in terms of CVs:

$$\mathrm{corr}(\hat{m}_i, \hat{m}_j) = \frac{\mathrm{Cov}(\hat{m}_i, \hat{m}_j)}{E[\hat{m}_i]E[\hat{m}_j]} \frac{1}{\mathrm{CV}_{\hat{m}_i}\mathrm{CV}_{\hat{m}_j}} = \frac{\widetilde{C}(\hat{m}_i, \hat{m}_j)}{\mathrm{CV}_{\hat{m}_i}\mathrm{CV}_{\hat{m}_j}},$$

where $\widetilde{C}$ is the normalized covariance. The connection between $\mathrm{corr}(\hat{m}_i, \hat{m}_j)$ and $\mathrm{corr}(n_i, n_j)$ becomes apparent once we realize that

$$\widetilde{C}(\hat{m}_i, \hat{m}_j) = (1 + CV_{1/N}^2)\widetilde{C}(n_i, n_j) + CV_{1/N}^2\,, \tag{S10}$$

which follows from the fact that $E[\hat{m}_i\hat{m}_j] = E[\beta]^2 E[n_i n_j] E[R^2]$. We are reminded that normalization by a noisy quantity (in this case $1/N$) can spuriously inflate positive covariances, and eliminate weak negative covariances (or inflate them if $\widetilde{C}(n_i, n_j) < -1$). From Eq. (S10) it follows that

$$\mathrm{corr}(\hat{m}_i, \hat{m}_j) = \mathrm{corr}(n_i, n_j)\frac{\mathrm{CV}_{n_i}\mathrm{CV}_{n_j}}{\mathrm{CV}_{\hat{m}_i}\mathrm{CV}_{\hat{m}_j}}(1 + CV_{1/N}^2) + \frac{CV_{1/N}^2}{\mathrm{CV}_{\hat{m}_i}\mathrm{CV}_{\hat{m}_j}}. \tag{S11}$$

To develop an intuition for the effects of sampling on gene-gene correlations, we assume that the variability between droplets in total counts $\mathrm{CV}_{1/N}$ is small, as is the case for undifferentiated ES cells. Then, using Eq. (S4) to relate $\mathrm{CV}_{n_{i,j}}$ to $\mathrm{CV}_{\hat{m}_{i,j}}$, Eq. (S11) becomes,

$$\begin{aligned} \mathrm{corr}(\hat{m}_i, \hat{m}_j) &= \mathrm{corr}(n_i, n_j)\alpha_i\alpha_j\,, \\ \alpha_{k\in\{i,j\}} &= \sqrt{\left(1 - \frac{1 + \mathrm{CV}_M^2 - E[\beta]}{F_{\hat{m}_k}}\right)} \end{aligned} \tag{S12}$$

where $F_{\hat{m}} = \mathsf{Var}(\hat{m})/E[\hat{m}]$ is the expected value of the *observed* gene Fano factor. To obtain Eq. (S12), we make use of the relationship

$$\frac{\mathrm{CV}_n}{\mathrm{CV}_{\hat{m}}} = \sqrt{\frac{\mathsf{Var}(n)}{E[n]}\frac{E[\hat{m}]}{\mathsf{Var}(\hat{m})}\frac{E[\hat{m}]}{E[n]}} = \sqrt{\frac{F_n}{F_{\hat{m}}}E[\beta]E[R]},$$

and then relate F_n and $F_{\hat{m}}$ by multiplying Eq. (S4) by $E[\hat{m}]$.

Note that the degree to which technical noise dampens the correlation between genes i and j is sensitive to the mean expression levels of both genes and to the sampling efficiency through the Fano factors. Since sampling efficiency is low, $E[\beta] \ll 1$, and we can approximate Eq. (S12) in terms of observable quantities only as

$$\mathrm{corr}(\hat{m_i}, \hat{m_j}) \approx \mathrm{corr}(n_i, n_j)\sqrt{\left[1 - F_{\hat{m}_i}^{-1}(1 + \mathrm{CV}_M^2)\right]\left[1 - F_{\hat{m}_j}^{-1}(1 + \mathrm{CV}_M^2)\right]},$$

giving Eq. (3) in Fig. 2G of the main text.

F.3 Identifying highly variable genes

A key goal of our data analysis is to identify genes whose expression in a population of cells is highly variable. More precisely, we wish to identify genes whose abundances are significantly over-dispersed relative to a Poisson distribution, which would result from uniform, non-fluctuating expression of transcripts in all cells. In this analysis, we use a test statistic that, at any given mean gene expression level, gives more weight to genes whose CV is many times larger than that of a Poisson random variable with the same mean. Based on Eq. (S4), a reasonable proposal for a test statistic, v, is:

$$v = \frac{\mathrm{CV}_{\hat{m}}^2}{\left(1 + \mathrm{CV}_M^2\right)\left(1 + \mathrm{CV}_{1/N}^2\right) \Big/ E[\hat{m}] + \mathrm{CV}_{1/N}^2} \qquad \text{(S13)}$$

By defining v in this way, we make concrete precisely what we do when we identify outliers by eye on a plot of genes' CV versus mean abundance such as Fig. 2F. The additive constant noise term $\mathrm{CV}_{1/N}^2$ keeps us from identifying a gene as highly variable in a population of cells if we can attribute much of that variability to differences in cell size. We infer $\mathrm{CV}_{1/N}^2$ from the data; for the ES cell data, $\mathrm{CV}_{1/N}$ ranges from $\sim 20\%$ on Day 0 to $\sim 35\%$ on Day 7 post-Lif withdrawal. For our RNA controls $\mathrm{CV}_{1/N}$ is much smaller – typically on the order of 5%, consistent with $\mathrm{CV}_{1/N}$ describing variability in total mRNA content per droplet. For both cells and RNA controls we calculate CV_M directly from the data; its values for the ES cell data are given in Table S1. The test statistic proposed here is similar to that proposed previously in (Brennecke et al., 2013), but with two key differences. First, here there is just one parameter to be inferred from the data $CV_{1/N}^2$, not two; second, we tested and found that the empirical distribution of v is not a χ^2 distribution as proposed in that study.

To develop a test for variability, we need a null distribution that describes the possible spread in v given that a gene's counts across cells are actually Poisson-distributed. For this purpose one may calculate v for a set of pure RNA controls, allowing for different values of $\mathrm{CV}_{1/N}$ and CV_M in each sample. One can then compute a p-value for each ES cell gene by comparing its v-score to the reference distribution, and thus test how many genes are significantly variable using Benjamini and Hochberg's method to control the false discovery rate (FDR).