Math 320 Research Project

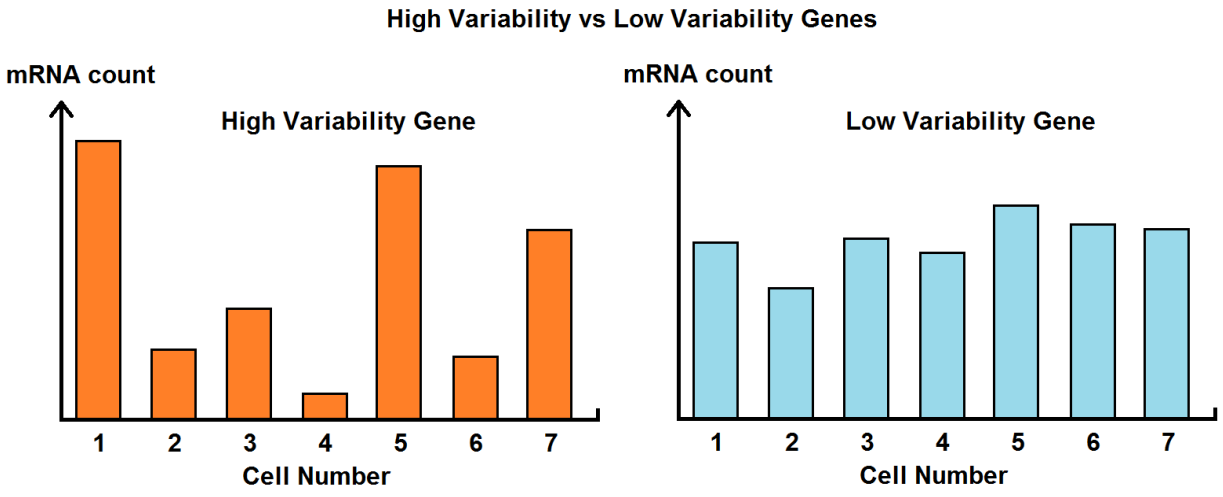
Report on “Accounting for Technical Noise in Single-Cell mRNA-Seq Experiments”

Jack Norleans and Yutong “Yolanda” Wang

Introduction

All living things are made of microscopic cells. The fundamental mechanism by which all cells are known to operate and survive involves three components: the cell’s DNA, mRNA, and proteins. DNA encodes the information needed to synthesize the thousands of different proteins a cell needs to perform its biological functions and survive. The exact sequence/information content of DNA will vary between different species. When a cell needs to create a protein, it will make a copy of the small section of DNA that codes for that particular protein. These small sections are called genes, and the small copies of these sections are a modified form of DNA known as mRNA. Molecular machinery in the cell directly convert or translate mRNA into the protein coded by the mRNA. To “express” a gene means to create mRNA for that gene and have that mRNA be translated into actual proteins by ribosomes. This well-established relationship between DNA, mRNA, and proteins is known as the Central Dogma and underpins all genetic research today, including the paper discussed in this research project. Essentially, the Central Dogma states DNA encodes numerous different genes, each of which codes for a unique protein. Ribosomes manufacture the protein based on the information from a gene, and mRNA molecules acts as the chemical messenger between DNA and ribosomes, delivering the information from a single gene on the DNA to the ribosome.

Most species’ DNA will contain tens of thousands of genes. Cells will express thousands of genes in different quantities to perform its biological functions and survive. For example, many genes will have undetectable expression, and other genes will be heavily expressed within a cell. A cell’s gene expression profile describes the extent to which each gene is expressed. Being able to accurately measure a cell’s protein or mRNA content and thus the cell’s gene expression profile is a powerful tool for understanding disease processes and normal biological functions. The algorithm in this paper aims to take quantitative data on the different mRNA sequences present in cell and find those genes that are expressed with highly variable expression. Some cells would heavily express these highly variable genes while other similar cells may barely express these genes.



Finding mRNA sequences (and thus their corresponding proteins) with high biological variability in expression will allow researchers to discover which proteins play critical roles in cellular processes. High biological variability is a likely hallmark of critical proteins because it means cells are adjusting their expression of these critical proteins to regulate larger scale biological functions. Altering their expression of critical proteins would allow cells to adapt to changing environmental conditions (e.g. diseased vs. healthy states). Thus proteins with high biological variability could serve as powerful drug targets for treating disease or valuable biomarkers for diagnosing disease in its early stages where medical treatments are most successful.

mRNA Sequence Counts

To generate the data used in this paper, a cell's mRNA content is reacted with special enzymes and tracer molecules to read the sequences of each mRNA molecule within a single cell. In the techniques described in this paper, the entire mRNA content of single cells are isolated, amplified, and indiscriminately sequenced in a high-throughput manner, producing over 30 million sequences. Prior to sequencing, numerous copies of each mRNA molecule are synthesized to increase the quantity of material available for the sequencing process to use. Amplified mRNA molecules from a single cell are then converted to a more chemically stable form prior to sequencing. Sequencing yields tens of millions of mRNA sequences derived from the mixture of amplified mRNA. The number of occurrences of each unique mRNA sequence are counted. It is assumed that mRNA sequences that are present in relatively larger quantities will be read more often, thus their sequences will appear more often in the data. The data the authors are working with are essentially tallies or counts of unique mRNA sequences detected within single cells.

This generates a large m-by-n array of data, where each element is a count of one unique mRNA sequence in one cell. Each row represents all the counts for a unique mRNA sequence, and the columns represent counts for all mRNA sequences within one individual cell. The authors' algorithm attempts to identify genes or unique mRNA sequences in this massive collection of sequencing data that are expressed with high biological variability between cells. A gene with high biological variability would be heavily expressed in some cells, but barely detectable in other similar cells.

mRNA is used to determine a cell's gene expression profile rather than protein concentration because mRNA can be sequenced and identified in a massively parallel manner regardless of the mRNA sequence. Proteins on the other hand, require a unique chemical assay (often an expensive specially designed antibody) for each protein. This makes it prohibitively expensive to identify and quantify the thousands of different proteins expressed in a cell.

Since each unique mRNA sequence is directly translated upon interaction with a ribosome, the quantity of any particular mRNA in a cell is directly related to how much of the associated protein the cell is producing. Each sequencing run may sequence short sections (50-100 base pairs) from over tens of millions of mRNA molecules. Each mRNA sequence from this process will produce a strand of bases (e.g. 5'-AACTGATTGTCTGGTATG…-3') that is overlaid on a genome for the organism the mRNA was taken from. For example, mRNA sequences from human cells will be matched to a human genome to determine which human genes the mRNA sequences came from. This matching process is very accurate since the probability of a mismatch occurring by chance, meaning the sequence is matched incorrectly to part of the genome that just happens to have the same sequence, is equal to:

$$probability\ of\ false\ positive\ match\ for\ each\ RNA\ sequence = length\ of\ genome * \frac{1}{4^n}$$

Where *n* equals the length or number of bases in the mRNA sequence. Many genomes are less than 2 billion bases in size, and the length of mRNA sequences obtained in this paper are around 50 base pairs. Assuming the genome of the organism of interest is 2 billion base pairs long (heavily overestimating the 135,000,000 base-pair genome length of *Arabidopsis thaliana* used in this paper) and the length of each mRNA sequence to be overlaid is 50 base pairs, the probability of a false positive match for each mRNA sequence comes out to:

$$P(false\ positive\ match) = \frac{2 \times 10^9}{4^{50}} = 1.578 \times 10^{-21}$$

The Problem the Paper Solves

The main challenge the authors face in detecting genes with high biological variability is that mRNA sequence count data for single cells is often extremely noisy. Single cells usually contain less than one nanogram of mRNA which, using modern sequencing technology, can only generate reproducible counts for only the most abundantly expressed genes. The "noise" associated with using small quantities of mRNA is tremendous for most genes in a cell. Between any two cells within a dataset, there can be significant variation in mRNA sequence counts for most genes even when the two cells express the exact same amount of the mRNA sequence. Thus it can be very difficult to distinguish true biological variation in mRNA expression from technique-related noise for most genes. The algorithm must be able to identify genes whose variability cannot be explained by noise related to the sequencing technique alone. The general strategy used by Brennecke et al. to accomplish this is essentially a hypothesis test conducted for every single gene counted by mRNA sequencing. For the data in this paper, mRNA sequence counts are taken from multiple cells. The variation for each gene is quantified as the variance in normalized mRNA sequence counts between all cells divided by the mean count squared. This gives the squared coefficient of variation for each gene across all the cells counted:

$$CV^2 = \frac{Variance}{Mean^2}$$

The strategy of the algorithm described in *Accounting for Technical Noise in Single-Cell mRNA-Seq Experiments* can be described as follows:

1. Count mRNA sequences
2. Normalize raw counts to account for different gross mRNA expression between cells
3. Calculate expected technical noise from reference mRNA counts
4. Hypothesis testing for each gene:
    a. Null hypothesis: gene's coefficient of variation $CV^2 \leq$ threshold variation. Threshold equal to expected technical variation plus a minimum biological variation threshold with a CV of 0.5 (or CV^2 of 0.25)
    b. Alternative hypothesis: gene's CV^2 is higher than threshold variation
5. Gene ontology (GO) analysis to connect each mRNA sequence to a cell function(s)

The first major step in Brennecke et al.'s strategy for solving the noise problem was to quantify the magnitude of random variations produced by the imperfect sequencing procedure alone (unrelated to

the tested cells' biology). This "technical" noise level is used as a benchmark against which experimental group mRNA sequence counts were compared to. The null hypothesis will be based on the expected technical noise quantified. If counts for any mRNA sequence in the experimental group produced higher between-cell variation than expected purely from technical variation/noise (rejects the null hypothesis), then that particular gene or mRNA sequence could be considered as having high biological variation. Brennecke et al. discovered that technical noise depended on the magnitude of the mRNA count for each gene (e.g. counts <100 produced high variation and counts >10,000 produced low variation), and created a regression formula to predict technical variation as a function of the magnitude of each gene's mRNA sequence count. The algorithm then calculates the chance or P-value a gene's measured variability or CV^2 is due to pure chance and the gene having zero true variability in expression. If a gene had a low P-value, it meant the gene was most likely truly expressed with detectable variation in single cells. The P-value however gives no indication of the extent of variability in the gene's expression.

The logic driving this strategy is that if a gene's variable expression cannot be explained by noise from the mRNA sequencing protocol alone, then that gene must have highly variable expression in cells due to real biological factors.

By using this strategy, the algorithm was shown to effectively identify, with an adjustable false discovery rate, genes that are actually expressed with high biological variability. Brennecke et al. tested the strategy against six *Arabidopsis thaliana* quiescent center of root cells (QC), seven non-hair root epidermis cells (GL-2), and 91 *Mus musculus* immune cells. Reference mRNA came from HeLa (clones of a famous human cervical cancer cell line) and an ERCC (External mRNA Controls Consortium)-formulated mRNA mixture specially designed to be used as a reference sample in mRNA experiments.

Normalization

Before any analysis can be performed, each individual cell's raw mRNA sequence counts for each gene must first be normalized to eliminate the effects of systematic errors from cell-to-cell variation in gross mRNA expression. Some single cells will produce significantly more or less mRNA overall than others due to differences in cell size. These differences in general mRNA expression are not related to the intrinsic properties of any particular gene but will inflate the level of technical noise if ignored. Thus, raw counts need to be normalized to a cell's total mRNA expression to reduce the impact of systematic cell-to-cell variation on the technical noise level. The actual normalization procedure divides a cell's mRNA counts by a size factor. To calculate the size factor, the geometric mean is taken for counts of each gene across all single cells used.

$$Geometric\ mean = (\prod_{i=1}^{n} k_i)^{1/n}$$

Each cell's actual count for that gene is divided by the geometric mean to obtain a size coefficients. Cells that systematically produce less mRNA and smaller mRNA counts across the board will produce small size coefficients under one. Cells that produce more mRNA than most will thus produce large size coefficients greater than one. There should be as many size coefficients as there are mRNA counts which is equal to the number of cells multiplied by the number of genes evaluated. The size factor for each single cell is equal to the median size coefficient for all the genes for which a geometric mean of counts could be obtained (if any cell had a count of zero for a gene, that gene would not produce a

geometric mean). All counts for each single cell are divided by the cell's size factor to obtain normalized mRNA counts.

Example of Normalization Process

	Cell Number							**Geometric mean**	**CV^2**
	1	2	3	4	5	6	7		
Raw Counts									
Gene 1	251	974	130	694	1197	160	80	319.242	0.833
Gene 2	111	278	76	278	242	1	1	39.785	0.772
Gene 3	801	1499	2500	1152	1781	450	675	1093.599	0.321
Gene 4	27	825	284	783	148	161	254	225.856	0.805
Size Coefficients									
Gene 1	0.786	3.051	0.407	2.174	3.750	0.501	0.251		
Gene 2	2.790	6.988	1.910	6.988	6.083	0.025	0.025		
Gene 3	0.732	1.371	2.286	1.053	1.629	0.411	0.617		
Gene 4	0.120	3.653	1.257	3.467	0.655	0.713	1.125		
Normalized Counts (Raw Count divided by median of Size Coefficients)									**CV^2**
Gene 1	330.550	290.584	82.078	246.068	445.141	350.619	184.370		0.185
Gene 2	146.179	82.939	47.984	98.569	89.995	2.191	2.305		0.621
Gene 3	1054.862	447.212	1578.431	408.459	662.319	986.115	1555.619		0.256
Gene 4	35.557	246.131	179.310	277.624	55.038	352.810	585.374		0.579

The above table is constructed from a small sample of data from mRNA sequence counts for GL-2 cells. After normalization, the CV^2 is reduced from those of raw counts for all genes. This indicates the normalization procedure can reduce the impact of systematic cell-to-cell variation in total mRNA expression on the magnitude of technical noise. Since systematic cell-to-cell variation does not actually contribute to true biological variability of a gene, it is important to account for its effects on technical noise.

According to the authors, the geometric mean was used instead of the arithmetic mean for normalization because the arithmetic mean was not as effective at normalizing mRNA sequence counts. This is most likely because the arithmetic mean failed to capture the "true expected count" for each gene. In this dataset, cells that systematically produced higher or lower counts for all genes often differ from other cells' counts by a factor of 2 or greater. This means an arithmetic mean would often produce a value that was heavily skewed from the main body or majority of mRNA sequence counts for any particular gene.

A theoretical explanation for why the geometric mean produced more effective normalization factors is that the mRNA sequence data tended to have a few extreme outlier counts for each gene. A high outlier could easily be more than double the median value observed for each gene/RNA sequence. Geometric means can "absorb" these high outlier so that have a smaller effect on the final calculated mean whereas an arithmetic mean would be more heavily skewed by a high outlier. The only drawback of using the strict definition of the geometric mean for determining normalization factors is that if one cell

in the dataset registered no counts (a count of 0) for an mRNA sequence, then the geometric mean for that mRNA sequence cannot be calculated. This means only mRNA sequences with high counts are used to calculate the normalization factor.

Spike-in Data and Determination of Technical Noise

To find the magnitude of technical noise produced by the sequencing procedure, a homogeneous mixture of known reference mRNA sequences is added or "spiked-in" as a "tracer" to the mRNA mixtures taken from single cells. Spike-in mRNA is sequenced alongside the sample mRNA in order to accurately quantify the magnitude of technical noise from the sequencing technique. Counts from reference mRNA sequences can be readily identified because the reference mRNA comes from a different species than the sample mRNA, thus reference mRNA will fail to map to the sample species' genome but will successfully map to a genome of the species the reference mRNA is derived from.

Each sample contains the exact same proportions of reference mRNA sequences. Thus any variation observed in the reference mRNA must be due to technical noise from the sequencing procedure. The authors can then quantify the expected technical noise based on CV^2 values from reference mRNA counts. mRNA counts from reference mRNA genes should theoretically produce zero variance if the sequencing procedure was perfectly accurate. In reality, there will be some variance observed even in the reference mRNA sequences due to the random nature of the chemical process used to sequence mRNA. Even then, the variance in reference mRNA gene counts should represent the minimum observable variance in the dataset. The non-reference mRNA counts from single cells will automatically have higher variances than reference mRNA counts since biological differences will produce additional variance for non-reference genes in addition to technical variation from the randomness of the sequencing procedure itself.

Based on the fact that the reference mRNA counts represent the minimum observable variance in counts between single cells, the increased variance for certain genes in the non-reference or experimental group mRNA counts above expected technical variance can indicate how likely that gene truly has highly variable expression. The next logical step in developing the algorithm would be to take the reference mRNA counts and derive a straightforward benchmark formula or values to represent the expected level of technical noise present in the dataset. Sample mRNA counts can be compared against this expected technical noise formula.

Technical Noise is Negatively Correlated with the Amount of mRNA Used in the Procedure

Technical noise would understandably come from the random nature of the chemical process that produces the mRNA sequence data. Some general formula would be needed to estimate technical noise as a function of some parameter of the mRNA sequence count data. Brennecke et al. soon discovered that technical noise in counts for any gene were heavily related to the abundance of its mRNA sequence in the sequencing process. This led the authors to derive a benchmark formula describing technical noise as a function of the magnitude of each gene's mRNA count. Technical noise in this paper is defined as the squared coefficient of variance (CV^2) defined by the equation:

$$CV^2 = \frac{Variance}{Mean^2}$$

CV^2 essentially is the squared root mean square of a dataset's average squared percent deviance from its mean value.

$$Variance = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

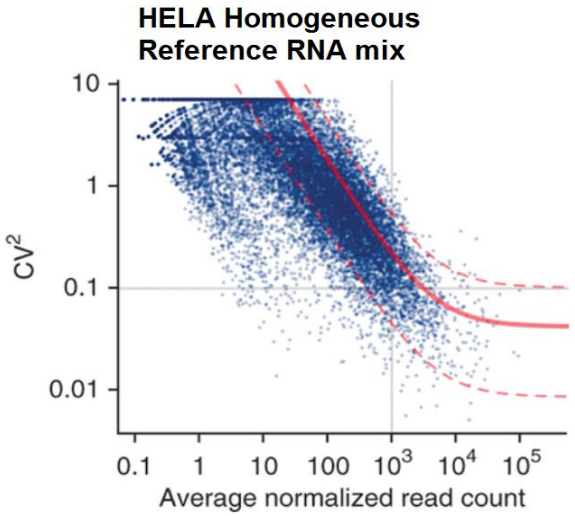
$$\frac{Variance}{Mean^2} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu}{\mu}\right)^2$$

$$Percent\ Difference\ from\ Mean = \frac{x_i - \mu}{\mu}$$

Where $\mu = mean$ of a set of data represented by x_i where i = [1,N] and N = # of elements in x_i,

When variance is converted into terms of percent difference rather than absolute difference, a sensible pattern is observed in the mRNA sequence count data. As shown in figures below taken from Brennecke et al., CV^2 for any particular mRNA sequence decreases with increasing size of its mRNA count. It appears that higher mRNA counts give more resolving power. The solid red line is the plot of a regression fit of CV^2 values for all genes with mRNA counts exceeding 500 (genes with less than 500 counts were considered too variable to be reliably included in the regression calculated). Formula **1** describes the general formula of the fit and provides a straightforward benchmark against which other data can be compared to. These counts below came from aliquots of homogeneous mixture of reference RNA sequenced in parallel which theoretically should produce zero CV^2 thus the figure below represents the minimum possible distribution of CV^2 values attainable with the sequencing procedure used here.

Another way to interpret the use of CV^2 rather than variance is to think of CV^2 as variance normalized for the mean. In physical systems, high counts are expected to produce high absolute variance and vice versa, making variance by itself an inaccurate measure of the tendency of a dataset to vary from its mean. Variance is biased by the size of the mean. To normalize, it is very reasonable to divide variance by the mean. Since the formula for variance involves squaring the difference between the mean and each value used to calculate the mean, the normalization factor must also be some kind of mean squared.



Formula for Estimating Technical Noise:

Expected $CV^2 = \alpha_1/\mu + \alpha_0$

(μ = Average normalized mRNA count)

This discovery seems very reasonable based on the fact that all steps of the chemical process used in Brennecke et al. involved enzymes randomly impacting mRNA molecules in solution and copying or sequencing the mRNA upon impact and complexion with the mRNA molecule. This means if particular mRNA sequences are only present in tiny amounts relative to the more abundant mRNA sequences, the chemical process may entirely "miss" these mRNA sequences. This possibility could occur at any step of the sequencing process, essentially ensuring that low abundance mRNA sequences will produce highly variable counts. Since the impact between mRNA and enzymes occurs randomly in each step, two identical solutions of mRNA may produce slightly different mRNA sequence counts. The random nature of gene-sequencing chemical process ensures this technical noise will always be present in the data.

The section of Brennecke et al. establishes that the precision or reproducibility of mRNA sequence counts is directly related to the quantity of mRNA used in the procedure and the actual magnitude of each count. The authors tested their procedure against 10, 50, 500, and 5,000 picogram (pg) samples of mRNA and found that their procedure generally gave highly reproducible counts with 5,000 pg samples, moderately reproducible counts at 500 pg, slightly worse reproducibility at 50 pg, and the worst reproducibility at 10 pg. At all quantities of mRNA used, the reproducibility was best when an mRNA sequence produced a high count (>10^5 counts) and was worst when low counts were used. Even with 5,000 pg of mRNA, sequences the produced less than 10 counts had almost no reproducibility. The quantity of mRNA used essentially determined the minimum count needed for an mRNA sequence to have good reproducibility, meaning repeated counts of the same sequence did not differ by more than 20% of each other. With 5,000 pg mRNA, even counts as small as 500 would more often meet this threshold than not. With 500 pg, counts would need to exceed 2,000 to be reproducible. At 50 pg, the reproducibility threshold approached 10^5 counts. At 10 pg, only counts high than 10^6 could exhibit the same reproducibility as 500-level counts at 5,000 pg, a 2,000-fold difference in the threshold count needed to maintain similar reproducibility.

Reproducibility essentially dictated the magnitude of technical noise that would be observed in actual experiments. The reproducibility test shows technical noise for any particular mRNA sequence count will mostly be a function of the quantity of mRNA used and the actual count number obtained for that sequence. This information heavily influenced the authors' decision in later experiments with *Arabidopsis thaliana* and *Mus musculus* cells to plot and fit equations to their data that described count variation as a function of its mRNA count number.

Calculation of Expected Technical Noise

Calculation for the expected value of technical noise is as follows.

When the authors scatter-plotted the CV^2 values of all their mRNA counts as a function of their mean count values, they noticed their plotted points fit very well to the formula:

$$\mathrm{Var}(Q_{ij}) = \widetilde{\alpha_1}\mu_i + \alpha_0{\mu_i}^2$$

From this observation, they derive a more rigorous equation that describes their expected technical noise for their mRNA counts.

They start from the assumption that the observed mRNA counts, K_{ij}, for any gene follows a Poisson distribution. The expected count value for each gene, λ, is directly related to the true concentration of each unique mRNA sequence, Q_{ij}, by the equation $\lambda = s_j * Q_{ij}$. s_j equals the size factor and is the proportionality constant that relates mRNA concentration Q_{ij} to expected mRNA count λ. The assumption that the observed counts K_{ij} follows a Poisson distribution for any gene is the main driver for deriving their equation for expected technical noise.[1]

$$K_{ij}|Q_{ij} \sim \text{Poisson}(s_j * Q_{ij})$$

$$(E(Q_{ij}) = \mu_i).$$

When quantifying technical noise, only counts from spike-in reference mRNA are used. In this case μ_i quantifies the number of instances of mRNA sequence *i* present in the mixture of spike-in RNAs added to the sample.

A Poisson process means that the distribution of the number of reads in any RNA assay can be modeled by the expression Poisson ($s_j * Q_{ij}$). RNA assay size equals the number of single cells (experimental groups) we test during sequencing. With greater RNA assay size, we will measure more raw RNA material and hence more aggregate reads across all cells. For a Poisson process, we have two basic assumptions:

(1) Each single cell should produce the same exact mRNA counts for each gene.

(2) The counts for each gene in a given sequencing run on a single cell is independent of the counts from other cells.

Therefore, $E(K_{ij}|Q_{ij})=s_jQ_{ij}$, $V(K_{ij}|Q_{ij})=s_jQ_{ij.}$, where we denote by $E(K_{ij}|Q_{ij})$ that the function at random variable Q_{ij} whose value at $Q_{ij} = q_{ij}$ is $E(K_{ij}|Q_{ij}= q_{ij})$, where q_{ij} denotes a specific Q_{ij} with some i & j.

According to the property of conditional expectation,

$$E(K_{ij})=E[E(K_{ij}|Q_{ij})]$$

For each fixed sample from single cell *j*, the size factor is considered constant because it derives from the information of so many genes that we consider its sampling variance negligible.

Since Q_{ij} is a discrete variable, for some specific library *j*,

$$E(K_{ij}) = \sum_i E((K_{ij}|Q_{ij}= q_{ij})*P\{ Q_{ij}= q_{ij}\} = \sum_i E(K_{ij}|Q_{ij})* P\{ Q_{ij}\} = \sum_i s_j Q_{ij} * P\{ Q_{ij}\} = s_j * \sum_i [Q_{ij} * P\{ Q_{ij}\}] = s_j * E(Q_{ij}) = s_j * \mu_i \quad (1)$$

$$Var(K_{ij}) = E[Var(K_{ij}|Q_{ij})]+Var[E(K_{ij}|Q_{ij})]= E(s_jQ_{ij}) + Var(s_jQ_{ij}) = s_j * E(Q_{ij})+s_j^2 * Var(Q_{ij})$$

$$= s_j * (1 + s_j\widetilde{\alpha_1})* \mu_i + {s_j}^2\alpha_0{\mu_i}^2 \quad (2)$$

$E({K_{ij}}^2)$= Var(K_{ij}) + $[\mathrm{E}(K_{ij})]^2 = s_j * (1 + s_j\widetilde{\alpha_1})* \mu_i + {s_j}^2(\alpha_0 + 1) * {\mu_i}^2$

$E(\widehat{\mu_i}) = E(\frac{1}{m}\sum_{j=1}^{m}\frac{K_{ij}}{s_j}) = \frac{1}{m} * \sum_{j=1}^{m}\frac{1}{s_j}E(K_{ij}) = \frac{1}{m} * \sum_{j=1}^{m}\mu_i = \mu_i$

Sample mean: $\widehat{\mu_i}=\frac{1}{m}\sum_{j=1}^{m}\frac{K_{ij}}{s_j}$ (3)

Sample variance: $\widehat{W_i} = \frac{1}{m-1}\sum_{j=1}^{m}(\frac{K_{ij}}{s_j} - \widehat{\mu_i})^2$ (4)

Combine equations (1), (2), (3) and (4), the expected value for sample variance is:

$$E(\widehat{W_i}) = E(\frac{1}{m-1} * \sum_{j=1}^{m}(\frac{K_{ij}}{s_j} - \widehat{\mu_i})^2) = \frac{1}{m-1} * \sum_{j=1}^{m}E[(\frac{K_{ij}}{s_j} - \widehat{\mu_i})^2]$$

$$= \frac{1}{m-1} * \{\sum_{j=1}^{m}Var\left(\frac{K_{ij}}{s_j} - \widehat{\mu_i}\right) + \sum_{j=1}^{m}[E\left(\frac{K_{ij}}{s_j} - \widehat{\mu_i}\right)]^2\},$$

in which:

$$\sum_{j=1}^{m}Var\left(\frac{K_{ij}}{s_j} - \widehat{\mu_i}\right) = \sum_{j=1}^{m}Var\left(\frac{K_{ij}}{s_j} - \frac{1}{m}\sum_{r=1}^{m}\frac{K_{ir}}{s_r}\right) = \sum_{j=1}^{m}Var\left(\frac{m-1}{m} * \frac{K_{ij}}{s_j} - \frac{1}{m}\sum_{\substack{r=1\\r\neq j}}^{m}\frac{K_{ir}}{s_r}\right)$$

$$= \sum_{j=1}^{m}[\frac{(m-1)^2}{m^2 * {s_j}^2}Var(K_{ij}) + \frac{1}{m^2}Var(\sum_{\substack{r=1\\r\neq j}}^{m}\frac{K_{ir}}{s_r})] = \sum_{j=1}^{m}[\frac{(m-1)^2}{m^2 * {s_j}^2}Var(K_{ij}) + \frac{m-1}{m^2 * {s_j}^2}Var(K_{ij})]$$

$$= \sum_{j=1}^{m}\frac{(m-1)}{m * {s_j}^2}Var(K_{ij}) = \sum_{j=1}^{m}\frac{(m-1)}{m * {s_j}^2}\left[s_j * \left(1 + s_j\widetilde{\alpha_1}\right) * \mu_i + {s_j}^2\alpha_0{\mu_i}^2\right]$$

$$= \frac{(m-1)}{m} * \sum_{j=1}^{m}\left[\frac{\mu_i}{s_j} + \widetilde{\alpha_1} * \mu_i + \alpha_0{\mu_i}^2\right] = \frac{(m-1)}{m} * [(\sum_{j=1}^{m}\frac{\mu_i}{s_j}) + \mathrm{m} * \widetilde{\alpha_1} * \mu_i + \mathrm{m} * \alpha_0 * {\mu_i}^2]$$

$$= \frac{(m-1) * \mu_i}{m} * (\sum_{j=1}^{m}\frac{1}{s_j}) + (\mathrm{m} - 1) * \widetilde{\alpha_1} * \mu_i + (\mathrm{m} - 1) * \alpha_0 * {\mu_i}^2$$

$$= (m - 1) * (\mu_i * \Xi + \widetilde{\alpha_1} * \mu_i + \alpha_0 * {\mu_i}^2), \text{ with } \Xi = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{s_j} \quad (5)$$

$$\sum_{j=1}^{m}[E\left(\frac{K_{ij}}{s_j} - \widehat{\mu_i}\right)]^2] = \sum_{j=1}^{m}[E\left(\frac{K_{ij}}{s_j} - \widehat{\mu_i}\right)]^2 = \sum_{j=1}^{m}[E\left(\frac{K_{ij}}{s_j}\right) - E(\widehat{\mu_i})]^2$$

$$= \sum_{j=1}^{m}[\mu_i - E(\widehat{\mu_i})]^2 = 0 \quad (6)$$

Combine (5) and (6) to yield:

$$\mathrm{E}(\widehat{W_i}) = (\Xi + \widetilde{a_1})/\,\mu_i + \alpha_0 * {\mu_i}^2, \text{ with } \Xi = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{s_j} \qquad (7)$$

To regress $\widehat{W_i}$ on $\widehat{\mu_i}$,

$$\mathrm{E}({\widehat{\mu_i}}^2) = \mathrm{Var}(\widehat{\mu_i}) + [\mathrm{E}(\widehat{\mu_i})]^2 = \mathrm{Var}(\frac{1}{m}\sum_{j=1}^{m}\frac{K_{ij}}{s_j}) + {\mu_i}^2 = \frac{1}{m^2}\sum_{j=1}^{m} Var(\frac{K_{ij}}{s_j}) + {\mu_i}^2$$

$$= \frac{1}{m^2}\sum_{j=1}^{m}\frac{1}{{s_j}^2} Var(K_{ij}) + {\mu_i}^2 = \frac{1}{m^2}\sum_{j=1}^{m}\frac{1}{{s_j}^2}[s_j * (1 + s_j\widetilde{\alpha_1}) * \mu_i + {s_j}^2\alpha_0{\mu_i}^2] + {\mu_i}^2$$

$$= {\mu_i}^2 * (1+\frac{\alpha_0}{m}) + \mu_i * \frac{\Xi+\widetilde{a_1}}{m} \qquad (8)$$

With (8), equation (5) becomes

$$\mathrm{E}(\widehat{W_i}) = \frac{1}{1+\frac{\alpha_0}{m}} * \mathrm{E}[(\Xi+\widetilde{a_1}) * \widehat{\mu_i} + \alpha_0{\widehat{\mu_i}}^2], \text{ with } \Xi = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{s_j}$$

As α_0 is typically small, the authors would neglect the corrective factor in the front of $\mathrm{E}(\widehat{W_i})$.

Take $\widehat{w_i} = \widehat{W_i}/{\widehat{\mu_i}}^2$ as the plug-in estimator for squared coefficient of variation (CV^2), and the authors could get the expected value of technical noise as a function of sample mean of normalized reads:

$$\mathrm{E}(\widehat{w_i}) = (\Xi+\widetilde{a_1})/\widehat{\mu_i} + \alpha_0$$

One caveat to this equation is that genes with low mRNA counts were excluded from the fit calculations. The authors excluded genes with very low mean mRNA counts, often below 100. In these cases, the authors concluded it is impossible to determine the gene's biological variation from observed variation in mRNA count. The authors state the maximum CV^2 for any gene must be equal to the number of cells *m* tested in the experiment based on the assumption that CV^2 for a gene is maximized when all but one of the tested cells has normalized counts for that gene equal to zero. In this case, CV^2 must be less than the number of cells tested *m*. However in their data, CV^2 values of 10 were observed when only 7 cells were sequenced. The authors do not have an explanation for this observation but conclude that when CV^2 exceeds its theoretical maximum value, it is impossible to measure biological variation from the observed CV^2.

Chi-Squared Hypothesis Testing for Statistically Significant Biological Variation

After the expected technical noise has been calculated as CV^2 as a function of mean mRNA count, hypothesis testing is done to find genes that are most likely expressed with true high biological variation. The null hypothesis is set as: "A gene's true CV^2 is less than or equal to a threshold CV^2 set as the expected technical CV^2 plus a small biological CV^2 of 0.25." The small biological CV^2 of 0.25 is factored into the threshold CV^2 for the null hypothesis because the authors want to only test for genes with exceptionally high biological variation. They are not interested in genes with small albeit

detectable true biological variation. This is why the threshold CV^2 for the null hypothesis is inflated above expected technical noise.

Any gene that rejects the null hypothesis is expected to have high biological variability. The authors conduct a one-sided test for the hypothesis by assuming the distribution of CV^2 values for mRNA counts follows a chi-squared distribution around the expected technical noise. Sample variance is calculated with the following formula with $\widehat{\mu_i}^B$ being the sample mean of the normalized counts, $s_j{}^B$ equal to the biological size factor calculated to normalize K_{ij}:

$$\widehat{W_i^B} = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{K_{ij}}{s_j{}^B} - \widehat{\mu_i}^B\right)^2,$$

The authors calculate the threshold CV^2 in the following manner. The author introduces a function to describe the expected value of true sample variance assuming the null hypothesis, which combines expected technical noise and the introduced biological CV^2 which is equal to α_F.

$$\Omega(\alpha, \mu) = \frac{\mu(\Psi + a_1 * \Theta) + \mu^2 * \alpha_F}{1 + \frac{\alpha_F}{m}}$$

$$\text{with } \alpha_F = \alpha_0 + \alpha + \alpha_0 * \alpha,\ \Psi = \frac{1}{m}\sum_j \frac{1}{s_j{}^B}, \text{ and } \Theta = \frac{1}{m}\sum_j \frac{s_j}{s_j{}^B}$$

Thus, the expectation of the threshold is $\Omega(\alpha_{th}, \widehat{\mu_i}^B)$ so that $\alpha_F = \alpha_0 + \alpha_{th} + \alpha_0 * \alpha_{th}$. Moreover, we expect $\widehat{W_i^B}$ follows approximately chi-squared distribution with m-1 degrees of freedom, because of its quadratic components. However, since the normalized sample counts are not truly independently and identically standard normal distributed, the expected distribution is not quite rigorous and waiting to be revised. After calculating $\widehat{W_i}^B$ and the threshold CV^2 value from expected technical noise and the added biological CV^2 of 0.25, the p-value is calculated by:

$$p = 1 - p_{\chi^2{}_{m-1}}\left(\frac{(m-1) * \widehat{W_i}^B}{\Omega(a_{th}, \widehat{\mu_i}^B)}\right)$$

Where $p_{\chi^2{}_{m-1}}$ denotes the cumulative distribution function of the chi-squared distribution with m-1 degrees of freedom.

The hypothesis test is adjusted to have a 10% false discovery rate for greater sensitivity or statistical power. When programming R, there is a function named p.adjust that can be used to account for the false discovery rate using the Benjamini-Hochberg method denoted as “BH”.

With p-values calculated, genes with high biological variability can be identified with the set 10% false discovery rate. Any genes with a p-value below the threshold p-value would reject the null hypothesis and hence have a high probability of having true high biological variability. Two different threshold p-values were considered in this paper, with a threshold p-value below 10^{-5} as clearly significant and below 10^{-4} as maybe statistically significant. We use it to produce a table of genes with high biological variability.

Gene Ontology Gives Meaning to mRNA Sequence Counts

Once genes with high biological variability have been identified, it is important to know the gene's function in the cell to extract meaning from the hypothesis testing algorithm. By linking each highly variable gene to different biological functions, it is possible to know which biological functions are experiencing the most variable regulation in the cell. The exact sequence of the mRNA can be used to identify the gene it codes for. A large database of genes has been established by other researchers which catalogs numerous genes by their name, sequence, and function in the cell. After matching is complete, nearly all genes with high biological variation would be sorted into a category of cell function (e.g. DNA replication, membrane transport, etc.).

Conclusion

The algorithm described in Brennecke et al. provides a reasonable way to identify a handful few hundreds of genes with high biological variability out of tens of thousands of genes expressed in a cell. The authors already test their algorithm against actual single-cell mRNA sequence counts and show how it can extract high resolution data on single-cell function. This technique would be greatly beneficial for future studies on cell function *in vivo* in large complex organisms, where even small sections of the organism may contain a diverse set of different types of cells heterogeneously distributed throughout the tissue. The high resolution afforded by this technique could lead to a detailed single-cell understanding of tissue function.

The main drawbacks associated with the algorithm are that only genes with reasonably high expression can be meaningfully processed. Genes with counts below 1,000 frequently demonstrate too much technical noise for the algorithm to accurately quantify its biological noise. In the present version of this algorithm, over half of all sequences obtained from each sequencing run are dedicated to calculating expected technical noise. The authors specifically mention this drawback and aim to find a way to use less sequencing reads to quantify technical noise in each sample. The authors attempted to use a small 92-gene spike-in rather than a massive HeLa cell spike-in, but the small 92-gene spike-in was not able to accurately quantify technical noise. Even in its current form, the algorithm provides a useful tool for single-cell genetic studies.

References:

1. Brennecke, P., Anders, S., Kim, J.K., Kolodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., Heisler, M.G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*. 10. 1093-1095.

Contributions of Team Members

Explanation of the biological applications, normalization function, and approach to quantifying technical noise was written by Jack Norleans. Explanation of the derivation of the expected technical noise and the details of the hypothesis testing was solved and written by Yutong Wang.