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 all proteins a cell needs to perform its biological functions and survive. The exact sequence/information content of DNA will vary between different species. mRNA relays instructions from small sections of DNA to molecular machines called ribosomes that translate the mRNA into proteins. Proteins execute the actual chemical processes that encompass a cell’s exact functions. The small sections of DNA covered by a single mRNA molecule are called genes and code for individual proteins. 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 understanding of the 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 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 may have zero detectable expression, and some genes may be overwhelmingly 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.

**Why Researchers Perform mRNA Sequencing**

Many techniques for measuring a cell’s gene expression profile involve quantifying the many different mRNA molecules found in a cell. Each mRNA coding for a unique protein will have a unique sequence of building blocks that encode the instructions for making a protein. In the techniques described in this paper, the entire mRNA content of single cells are isolated and amplified, meaning several copies of each mRNA molecule are made to improve the signal-to-noise ratio. The mixture of thousands of different mRNA molecules from a single cell are converted to a more chemically stable form while conserving the mRNA’s information content. Tens of millions of random mRNA molecules in the mixture are sequenced, and repeated occurrences are tallied for each mRNA sequence among the millions of sequenced molecules. The total dataset containing counts for each particular mRNA sequence represents the gene expression profile for the cell or mRNA mixture. The paper we will discuss describes an algorithm that identifies variably expressed genes within an mRNA sequence count dataset whose variability cannot be explained by noise or other technical factors alone. By using this strategy, the algorithm can effectively identify variably expressed genes with adjustable false discovery rates.

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.

**The Problem the Paper Solves**

The challenge of identifying variably-expressed mRNA sequences/genes is that counts of mRNA sequences from single cells can be extremely noisy. Between any two cells within a dataset, there can be significant variation in mRNA sequence counts for most genes even when the two cells expressed the exact same amount of the mRNA sequence. Thus it can be very difficult to distinguish noise from true biological variation in mRNA expression for most genes.

The aim of *Accounting for Technical Noise in Single-Cell mRNA-Seq Experiments* is to design an algorithm in *R* that can identify particular mRNA sequences (and their respective genes) with statistically significant variability in expression within large datasets of mRNA sequence counts. 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 biological factors. Such genes with high biological variation are likely to be critical components for regulating cell physiology since the cells are significantly adjusting their expression to adapt to changing environmental conditions. 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) and survival needs. 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.

Brennecke et al.’s strategy for solving this problem was to quantify how much does random variations in the sequencing procedure (unrelated to the tested cells’ biology) contributes to the variation or noise produced in single-cell mRNA sequence counts. This “technical” noise level is used as a benchmark against which experimental group mRNA sequence counts were compared to. If counts for any mRNA sequence in the experimental group produced higher between-cell variation than expected purely from technical variation/noise, then that particular gene or mRNA sequence could be labelled 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. A statistical test is determined for the variance value of all genes’ count data to find highly variable genes. Any genes with statistically significantly higher variance than expected (based on the technical noise fit as a function of count magnitude) are considered to have high biological variance.

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 on between-cell variance. Some single cells will produce significantly more mRNA in general than others due to differences in size. To account for this, each cell’s mRNA counts are divided by a size factor. The geometric mean is taken for counts of each gene across all single cells used. 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.

[INSERT SAMPLE DATA TABLE]

According to the authors, the geometric mean was used instead of the arithmetic mean for normalization because the arithmetic mean was ineffective at normalizing counts across all cells (across the columns). 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 meant 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. The failure of the arithmetic mean to properly normalize the data can be observed in the following example data taken from the paper’s supplementary information section.

[INSERT EXAMPLE NORMALIZATION ANALYSIS]

This excerpt of data is taken from the mouse single-cell mRNA sequence counts. In the reference mRNA counts, which theoretically should have virtually zero variance in normalized counts between cells, using the arithmetic mean in the normalization algorithm results in high between-cell variance. The geometric mean on the other hand produces much lower variance. The authors noticed the geometric mean was much more effective at reducing noise due to systematic errors in their mRNA sequence counts and thus chose to use the geometric mean over the arithmetic mean in their normalization algorithm. 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 reference mRNA is added or “spiked-in” as a form of “tracer” to the mRNA mixtures taken from single cells. The reference mixture of mRNA from a different species is homogenized and combined in known proportions (e.g. 5 pg of reference mixed mRNA with 7 pg of the unknown cell’s mRNA) with experimental group mRNA from each cell and sequenced together. Reference mRNA sequences can be easily differentiated from non-reference mRNA sequences since reference mRNA comes from a different species than the non-reference mRNA. Thus, most of the reference mRNA sequences will fail to map to the genome of the species the non-reference single-cell mRNA is obtained from.

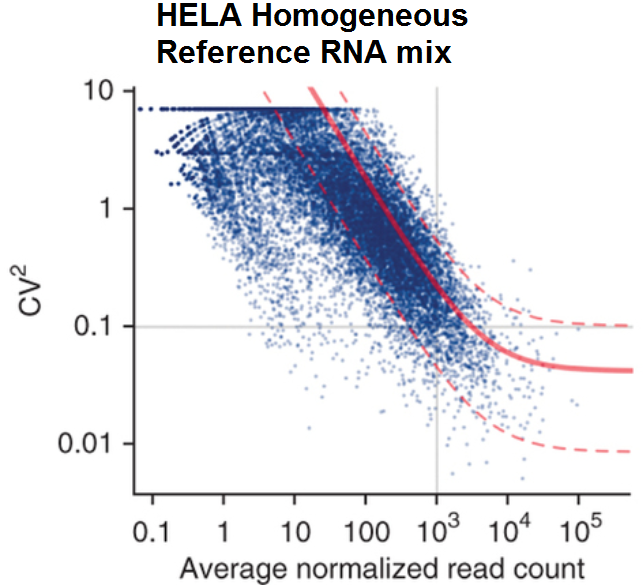
Since the reference mRNA is a homogeneous mixture, each single cell’s mRNA mixture should contain the exact same composition of reference mRNA mixture. Since each experimental group’s mRNA was “spiked” with the exact same mixture of reference mRNA, the between-single-cell variation in reference mRNA sequence counts must be only caused by the noisiness of the physical techniques used to prepare and sequence the cell’s mRNA. 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 genes 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 due to 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 level technical noise present in the dataset. Non-reference mRNA counts can be compared against this straightforward technical noise coefficient(s) or 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 cumulative variance value (CV2) defined by the equation:

CV2 essentially is a measurement of a sample’s tendency to produce variable counts normalized for the size of the count. High counts are expected to produce high absolute variance and vice versa, making normalization necessary. 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.

As shown in figures below taken from Brennecke et al., technical noise for any particular mRNA sequence decreases with increasing size of its mRNA count. The solid red line is the plot of a regression fit of CV2 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 CV2 thus the figure below represents the minimum possible distribution of CV2 values attainable with the sequencing procedure used here.

Benchmark Formula for Estimating Technical Noise:

***Approximate CV2***

**= *ã*1/*μ* + *α*0**

*(μ = Average normalized read 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 first figure in the paper 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 (>105 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 105 counts. At 10 pg, only counts high than 106 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 gene variance as a function of its mRNA count number.

**Chi-Squared Test for Statistically Significant Biological Variation**

After a technical noise baseline has been established, CV2 values are calculated for all genes in the single cell mRNA sequence counts. CV2 from single-cell genes are plotted against the magnitude of the average mRNA count for that gene. Each gene is tested based on its CV2 for whether it satisfies the null hypothesis that the gene’s true variance is at most equal to the technical variance expected for the magnitude of its average mRNA count plus a value equal to its mean squared multiplied by a coefficient of biological variation of 0.5 (essentially a CV value attributable to biological variation in gene expression). This combined value represents the expected total variance for a gene, given the level and technical noise and a reasonable amount of biological variation expected for any expressed gene. A one-sided test was conducted for the null hypothesis using a Chi-Squared function. Inputs of the expected total variance and the actual variance yielded a P-value. The P-value states the probability of observing the obtained total variance for a gene if there were no physical reasons for gene to have biological variability. It states the probability of achieving the observed total variance by pure chance assuming the null hypothesis is true. A low P-value indicates the null hypothesis is most likely false and should be rejected. No cutoff P-value was stated in the paper for rejecting the null hypothesis, but it can be set anywhere from 0.05 to as low as 0.00001. Rejection of the null hypothesis indicates the gene in question most likely has high true biological variability in its expression among the single cells assayed.

Application of this hypothesis test to each gene yielded a list of high-biological-variability genes. There is a false discovery rate associated with this test. For this paper, the authors report a ~10% false discovery for genes with high biological variability in expression.

**Gene Ontology Gives Meaning to Data on Biological Variability in mRNA Sequence Counts**

Once genes with high biological variability have been identified, it is important to know the gene’s function in the cell. By linking each highly variable gene to different biological functions, it is possible to sees which biological functions are experiencing the most changes or adaptations in the cell. All high-variability genes are categorized by a gene ontology algorithm. In this context, gene ontology essentially refers to the categorization of genes into various categories with each category representing a certain class of related biological functions (e.g. DNA replication, membrane transport, etc.).