Investigator: Dominic D. LaRoche
Proposal Title: Novel methodology for evaluation of next-generation sequencing measurements.

1 Significance

The accuracy and precision of many RNA measurement systems is largely unknown despite the importance of these measurements in modern biological research. We define an RNA measurement system as any sequence of procedures designed to quantify RNA from a biological sample and provide data for analysis. There are currently over 5 widely used RNA measurement systems on the market, each with a different form of technology to achieve the final measurement. Our proposed research will provide a methodology for studying measurement error in NGS-based RNA measurement systems. Specifically, we will provide a method to directly compare the precision of any two RNA measurement systems using only technical replicates from each system. Even though RNA is widely studied, the precision of measurements may limit the quality of inference and slow the rate of scientific progress. Continued progress in fields that rely on NGS-based RNA measurement systems will depend on understanding how and why errors occur.

There is currently no robust methodology for comparing measurement precision between NGS-based RNA measurement systems. The current approach to comparing measurement systems is to indirectly compare each system's reproducibility and repeatability, an approach with limited utility (Lovell et al. 2015). Alternatively, others have used the final outcome measure, i.e. differential expression of genes, as a measure of performance. However, the outcome of an analysis is only a crude measure of measurement precision at best, even when using simulated data where the truth is known. Our method will allow direct comparison of the precision of any two measurement systems. This will allow for: 1) evaluating measurement systems for diagnostic procedures; 2) improving a single measurement system through experimentation to determine where errors arise; 3) evaluating the relative impact on precision from the numerous data normalization methods currently in widespread use; and 4) selecting the best measurement system and normalization method for any given experiment. Finally, we will provide all of our methodology as a freely accessible web application.

2 Innovation

We will apply a little known mathematical framework, Relative Sensitivity, originally developed by John Mandel in 1985 for evaluating measurement in analytical chemistry. This method does not require knowledge of the true amount of the sequences being measured, a quantity that is often difficult or impossible to know in RNA measurement problems. Relative sensitivity provides the framework for making comparisons among normally distributed measurements. Our approach is conceptually innovative because we will need to derive the formulas necessary to apply relative sensitivity to less well-understood count data, often with multiplicative errors, arising naturally from NGS-based measurement systems.

A new methodology is only useful if it is accessible to a wide range of researchers and scientists. However, many advances in statistical methodology take considerable time for adoption by the general scientific community because of the barriers to implementation. Our approach is technically innovative because, if we find utility in our proposed methods, We will provide a software suite for implementing our methods through the recently developed web platform for R programs, shinyapps.io. This platform allows users to interact with statistical analyses pre-programmed in R through a web-based graphical user interface (GUI). The platform greatly reduces the barriers to method implementation because it does not require anything to be installed on user computers and can provide elaborate point-and-click documentation to guide the user through the analyses. We will provide our suite of methods as a dynamic GUI available for free to researchers who wish to compare measurement or normalization systems on their own data. This platform has never been used for widespread implementation of a new method but holds great potential as a translational statistics tool.

3 Specific Aim 1 Research Plan

3.1 Background and Rationale

There is no established method for directly comparing the precision and accuracy of NGS-based RNA measurement systems. The rapid adoption of NGS-based RNA measurement systems across a wide range of biological disciplines necessitates improved understanding of the precision of these systems. There are more than four different measurement systems which utilize next generation sequencing technology currently in widespread use. These systems all have different technological features and are unlikely to produce identically precise results. The ability to compare measurement systems would also enable scientists to select the optimal system for their experiment and allow manufacturers to identify sources of error and make improvements. Currently, measurement systems are evaluated on the basis of repeat-ability and reproducibility but there is no way to directly compare existing metrics of repeat-ability and reproducibility among competing measurement systems due to the differences in scales and underlying distributions of the measurements.

The objective of this aim is to establish a methodology to directly compare two measurement systems. We hypothesize that the method of relative sensitivity will provide a powerful and useful framework for making these comparisons. The theory of relative sensitivity was originally developed by John Mandel in 1984 as an extension of his original work on the sensitivity of analytical chemistry measurements (Mandel 1957). Relative sensitivity has several key properties which make it suitable for this application. For example, relative sensitivity is not affected by the scale of the measurement, which can vary substantially between NGS-based measurement systems. Moreover, most measurement systems include some monotone transformation of the final result for analysis, of which there are many choices, and relative sensitivity is invariant to these transformations.

We will use both simulated and real data from 4 measurement systems to evaluate the estimates of precision and accuracy we generate from the relative sensitivity framework. Simulated data is necessary as the truth of real-world measurements can never be known. We will focus on simulating the measurement process, rather than the final result as is typically done, so that we can understand the impacts of each step in the measurement process. We will confirm the utility of our method by evaluating the accuracy and precision of four competing measurement systems: HTG EdgeSeq, NanoString nCounter, Illumina RNASeq, and TaqMan Gene Expression assays. We will use technical replicate samples and quantify each sample replicate 24 timeson each platform. We have already completed sample quantification on HTG EdgeSeq and NanoString nCounter.

A metric for comparing NGS-based measurement systems will have immediate utility in identifying optimal measurement systems and sources of measurement error. This method will provide a critical window into the performance of measurement systems for both consumers of these systems and the manufacturers of these systems. Due to the diverse nature of the technology of systems currently in use, it is likely that the performance of these systems will correlate with elements of the technology utilized. We believe understanding these differences in performance will be important for mitigating problems at the development level and selecting the appropriate method for a given experiment.

At the completion of this aim we expect that we will provide an implemented methodology for comparison of any 2 NGS-based measurement systems such that a scientist with little understanding of the underlying theory of relative sensitivity will be able to implement the methodology and interpret the results.

3.2 Experimental Plan

The theory of relative sensitivity was initially presented in John Mandel's book *The Statistical Analysis of Experimental Data* (1984) but was never widely adopted in the analysis of measurement systems outside of analytical

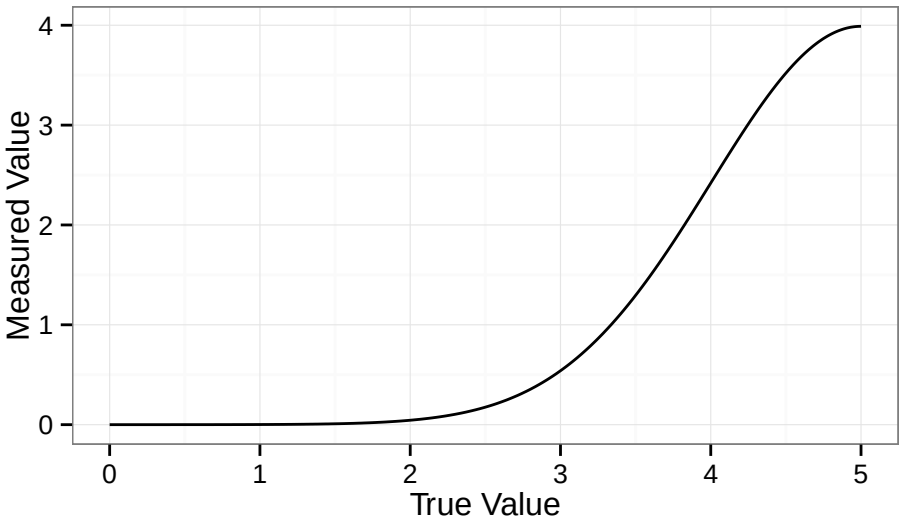

Figure 1: A single sensitivity curve for a measurement system representing the relationship $Y = f(X)$, where $f(X)$ is some unknown non-linear function of X. The sensitivity of the measurement is greatest between the true values of 3 and 4.5 and smallest at 0 and 5. This measurement system would not be a good choice for those interested in measuring true values of less than 2 since the method is unable to differentiate between these values.

chemistry (with few references even within analytical chemistry). We attribute this to the relatively spotty uptake of statistical procedures in general and to the limited utility of such a method when calibration curves are more easily attainable. However, we believe the theory of relative sensitivity provides a simple, yet powerful, statistical framework for the evaluation of complicated NGS-based measurement systems.

The theory of relative sensitivity is an extension of Mandel's work on estimating sensitivity curves (1957) which removes the need to know the actual analyte concentration in order to evaluate the precision of a measurement. The sensitivity of a measurement is the slope of the functional relationship between a property of interest and its measurement, $Y = f(X)$, where Y is the measured value, X is the true value, and $f(X)$ represents some unknown relationship between these values. The function $f(X)$ maps a can be non-linear and/or non-increasing (or decreasing) in practice but must be monotonic to be useful. If the true count of target RNA is known for each sample and a variety of samples are measured, each with a different amount of target RNA, then a calibration curve can be constructed as in figure 1.

For a given measurement system, steep slopes ($f(X)$) correspond to greater sensitivity because a steep slope results in large differences in the measured value for small differences in the corresponding property being measured (fig. 1). The error around a sensitivity curve for a measurement process also affects the utility of the measurement. For example, if the slope is small but there is little error then the measurement can still discriminate between different states of the property whereas large error can "swamp out" a steeper slope (fig. 2).

In order to construct and evaluate a sensitivity curve as described above one must know the true value of the property being measured. For many NGS-based measurement systems this property is either unknowable or is exceedingly difficult to know. However, relative sensitivity can be used to compare two measurement systems without knowing the true state of the underlying property being measured. A key insight of Mandel's was that two measurements of the same quantity must be related. Mandel used this insight to construct the relative sensitivity curve as a way to compare two measurement systems, resulting in measurements Y_1 and Y_2, without knowing the true value of the measured quantity. The relative sensitivity ratio takes the form:

$$RS(Y_1/Y_2) = \frac{|dY_1/dY_2|}{\sigma_{Y_1}/\sigma_{Y_2}}.$$

In this formulation, dY_1/dY_2 represents the slope of the relationship between Y_1 and Y_2, i.e. the slope of the

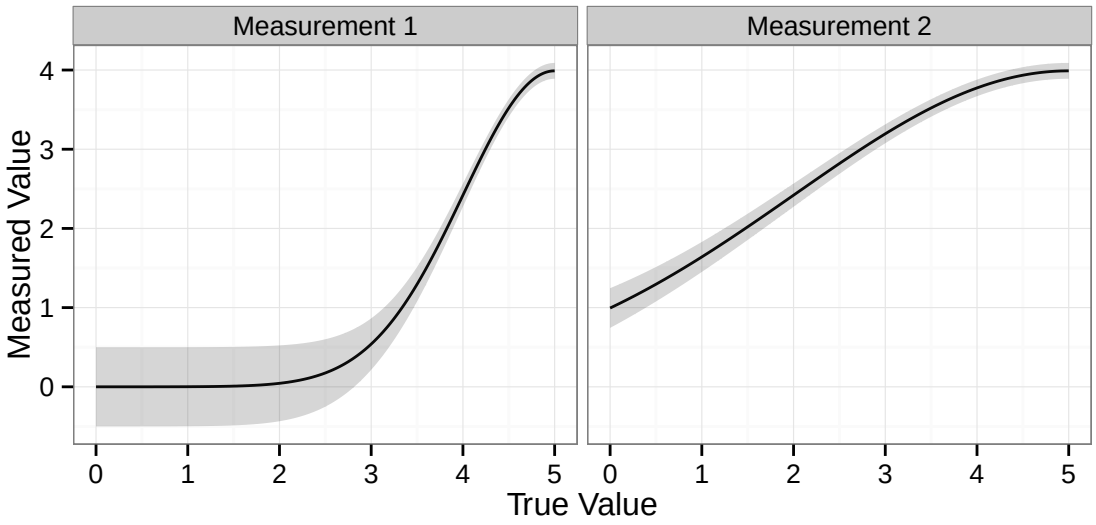

Figure 2: A comparison of sensitivity curves for 2 hypothetical measurement systems. Here measurement 2 is clearly superior to measurement 1 for true values less than 3, whereas measurement 1 is clearly superior to measurement 2 for values greater than 3.5. The choice of the optimal measurement method is dependent on the expected range of the true value being measured.

line in a plot of Y_1 versus Y_2, while σ_{Y_1} and σ_{Y_2} represent the standard deviations of the two measurement systems. This ratio does not involve the true value of the property being evaluated while providing a simple metric of relative utility for two measurements. For most measurement systems the sensitivity of the measurement is not linear (fig. 1) and, therefore, the slope of the relationship between measurements Y_1 and Y_2 will also vary (fig. 3). The resulting changes in the value of the relative sensitivity ratio produce the relative sensitivity curve (fig. 4). When the relative sensitivity is less than 1 method 2 is better whereas when the relative sensitivity is greater than 1 method 1 is better. From figure 4 we can see that method 2 outperforms method 1 until the true value of the measured quantity is about 3.25, at which point method 1 becomes preferable.

To estimate the relative sensitivity, one needs to know the slope of the relationship between the two measurement systems being compared as well as the standard deviations of the two measurements. The standard deviations of the two measurements can be easily estimated from technical replicates. We will compare two approaches for estimating the relationship between the two measurements. The original formulation by Mandel suggested directly estimating the slope of the relationship by regressing y_1 onto y_2. Unfortunately, this method assumes no measurement error in y_1 which is clearly not plausible. We will, instead, compare two different approaches to account for the measurement error in both measurements. First, for an assay with n probes we will use an empirically estimated slope by calculating the change in each measurement from each rank-ordered probe to the next higher probe, $\delta y_i = y_{ij} - y_{i(j-1)}$ where i =1, 2 and j =2, ..., n. We will then calculate the empirical slope, s_j as the ratio $s_j = \frac{\delta y_{1i}}{\delta y_{2i}}$. Second, we will use a model based Bayesian approach to simultaneously estimate the slope and measurement errors.

We will test the utility of our method by simulating the measurement processes under consideration. Frequently, investigators will simply simulate the final measurement using assumptions which favor the method under investigation (citation here). We will simulate the measurement process for each of the 4 measurement systems under consideration. This approach will have 3 benefits: 1) we believe the final simulated data set will more closely match real data likely to be encountered by end users, 2) we will be able to investigate how perturbations for a given step within a measurement system affect the final utility of the measurement, and 3) simulated data for each measurement protocol will have arisen from the same true concentration of target RNAs. We will also test our method using real data collected from all four measurement systems under review. We have already collected this data for identical samples on the HTG EdgeSeq and the NanoString Ncounter platforms.

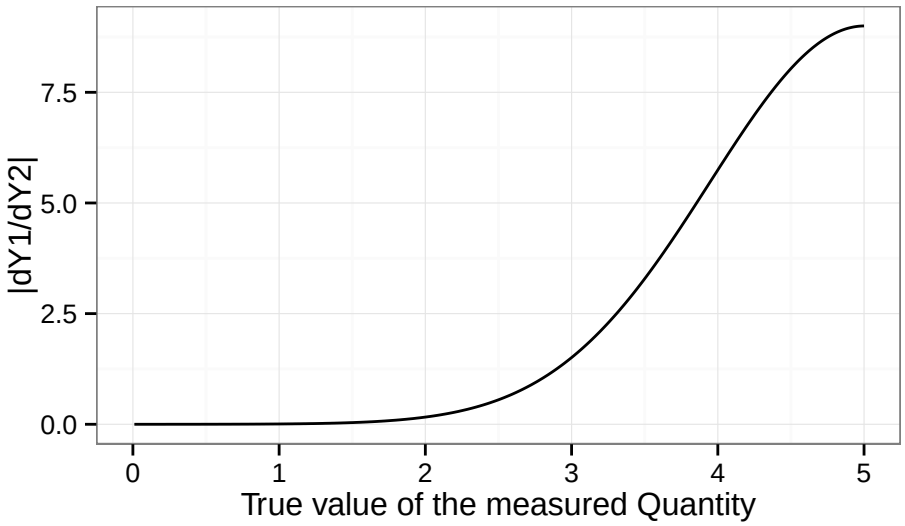

Figure 3: The slope of the relationship between the two measurement systems shown in figure 2, dY_1/dY_2, for each value of the true quantity being measures (not typically known).

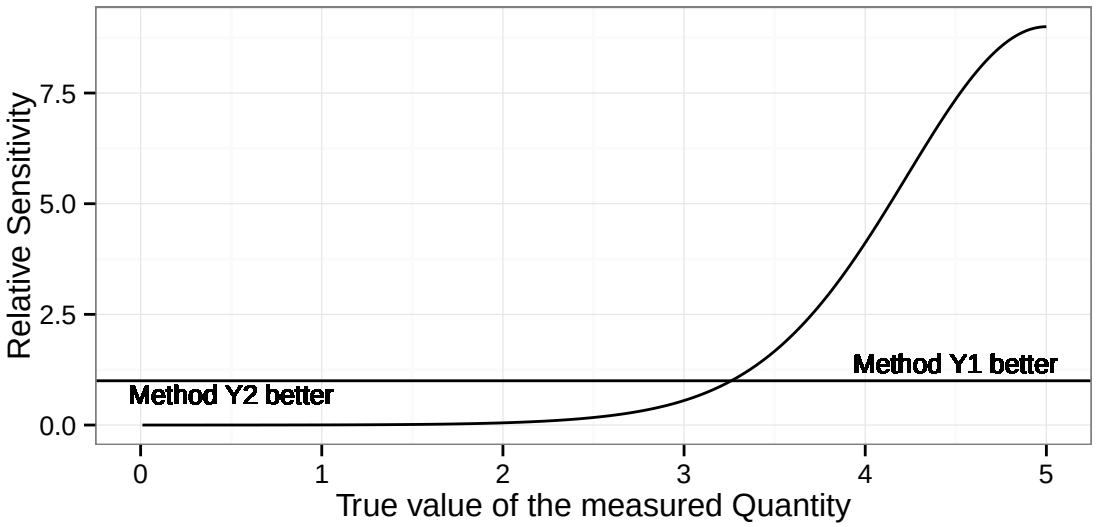

Figure 4: The relative sensitivity for the two measurement systems resulting in the measured quantities Y_1 and Y_2. Here we have the truth plotted along the x axis while in practice this would not be available in which case substituting either Y_1 or Y_2 would suffice. When relative sensitivity is < 1 method Y2 is better whereas when relative sensitivity is > 1 method 1 is better. The relative sensitivity shown here can be compared to the two measurements shown in figure 2.

We will implement the method in the open source statistical package R and provide a open source package of functions to import data and compare measurement systems. We will also provide a graphical user interface to this package by creating an interactive web page that allows users to point and click through an analysis. We will host the application through the free shiny.io service provided by Rstudio.

3.3 Expected Outcomes, Potential Problems, and Alternative Strategies

At the completion of this aim we expect to have an implemented methodology which can compare any two NGS-based measurement systems in terms of precision and accuracy. We expect to have our method freely available for use by scientists through a user-friendly graphical user interface available to anyone with an internet connection. We also believe the FDA will benefit from our methodology due to the increasing use of NGS-based measurement systems to make clinical decisions.

The utility of relative sensitivity to evaluate subtle differences in the performance of NGS-based measurement systems is obviously unknown and there are several challenges to this proposed methodology. Specifically, relative sensitivity requires that the two measurement systems are measuring the same quantity. We will used specially prepared universal reference RNA (Agilent Technologies) to ensure the highest similarity across samples run on different platforms. We will also ensure that assays developed for different platforms are actually targeting the same RNA sequence. Moreover, our estimation of the relationship between any two measurement systems requires a broad range of target RNA abundance. Our previous experience suggests that the combination of universal RNA samples and broad panel probes for each measurement system will fulfill this requirement. However, it is possible, even after taking these precautions, that we will be unable to differentiate between existing systems using our proposed estimation of the relative sensitivity curve. Given this scenario we will still be guaranteed to identify an upper bound on the difference in precision between any two measurements using information obtained through simulations.

If we are unable to detect a difference in precision between measurement systems using our proposed estimation scheme we will re-formulate our estimators based on the maximum-likelihood. These estimators will require more assumptions about the data but will likely yield smaller confidence intervals around our estimates of precision due to these assumptions. Although we believe our proposed method with fewer assumptions is more desirable, the assumptions made by the maximum likelihood estimates are common when dealing with NGS-based data so we feel this method would also provide utility to scientists and manufacturers.

4 Aim 2 Evaluation of frequently used normalization procedures.

4.1 Background and Rationale

Data from NGS-based RNA measurement systems must be normalized prior to most subsequent data analyses to remove systematic technical effects. There are currently over 7 methods to normalize NGS-based measurement data in widespread use and no clear guidance exists for selecting a method. A recent comparison of 7 popular normalization methods by Dillies et al. (2012) was ambiguous about the optimal normalization method and under what conditions optimality would hold. Two years after the comparison was published we still see widespread use of nearly all of the methods compared, indicating uncertainty among scientists and statisticians about the optimal method. At best, researchers are selecting methods based on limited, and sometimes conflicting, information. At worst, researchers are selecting a normalization method which they find to be optimal in terms of the results they expect (or would like) to see.

Our aim is to create an objective measure of the relative efficacy of the 7 most popular normalization methods so that researchers can select the optimal method for their data *without* basing this decision on the final outcome. The normalization procedure can be thought of as part of the measurement system. As such, we hypothesize

that relative sensitivity can be adapted to provide a relative measure of performance for any pair of normalization methods. By comparing the impact of each type of normalization on the relative precision of the measurement we will provide necessary information for researchers when selecting a normalization method.

The review article by Dillies et al. utilized both simulated and real data. For both types of data the authors used the number of differentially expressed genes detected after normalization as the measure of performance for each normalization method. For the simulated data this is an appropriate, if crude, measure of performance. However, for the real data since the number of truly differentially expressed probes is unknown this is not appropriate. The practice of selecting a normalization method based on 'optimal' final results could lead to an increase of false discoveries due to the circular nature of the decision. Moreover, RNA measurements are utilized for more than just differential expression analysis and no comparison of normalization methods has been conducted outside of the differential expression paradigm. Unfortunately, without some objective measure or clear guidance this practice will likely continue. Our research will provide an objective measure and guidance about the impacts of all 7 commonly used normalization methods. We believe this methodology will improve the quality of research across the wide range of research areas which utilize this type of data.

4.2 Experimental Plan

Existing normalization methods are based on different assumptions about the underlying nature of the observed data (lots of refs). We will simulate observed data under a variety of these assumptions in order to compare normalization methods under conditions when assumptions are and are not met. We will use relative sensitivity to evaluate each normalization method against each simulated data type to determine when each method is either appropriate or inappropriate and optimal or sub-optimal. We will define a method as being appropriate when the inferences that would be made from normalized data (e.g. differential expression results) are consistent with the truth. We will define an optimal method as the method normalization which produces final measurements with the highest precision across the entire range of measurement.

Some methods may outperform others in only a particular range of measurement. We will divide the range of measurement into high, medium, and low expressed probes by dividing the simulated measurements into quadrilles. Low expressed probes will be defined as those less than the first quartile, medium expressed as those between the 1st and 3rd quartiles, and high expressed as those above the 3rd quartiles. Since different normalization methods are likely to place different sets of probes into each of these categories we will identify which probe is assigned to each category based on the random variable, λ, associated with the expected count for that simulated probe.

Simulations provide important information about the performance of normalization under tightly controlled conditions where the truth is known. However, real data often do not conform with the simplified and controlled nature of simulations. Therefore, we will also compare normalizations on real data using the technical replicate samples quantified on each of the 4 platforms identified in aim 1. We will use relative sensitivity to evaluate the effect of each normalization method for the data generated by each platform.

4.3 Expected Outcomes, Potential Problems, and Alternative Strategies

Previous research has shown many normalization methods to perform similarly on real observed data (Dillies et al. 2012; Rapaport et al. 2013). Therefore, we expect that the performance, as measured by relative sensitivity, of each normalization method to also be quite similar. However, we believe relative sensitivity will provide much more information about the precise nature of the differences in normalization methods than previous comparisons. Previous comparisons of normalization methods relied on using the outcome, e.g. the number of differentially expressed probes, as the measure of the effect. The outcome is a very crude measure of the effect of a normalization procedure and has been unable to differentiate among competing procedures in previous research. Relative sensitivity has the potential to directly compare the effects of each normalization method allowing the

direct comparison of each method with respect to any, or all, measurement range(s). We believe this level of detail will enable us to differentiate the impacts of closely related normalization methods such as TMM (Robinson and Oshlack 2010) and Quantile (Amaratunga and Cabera 2001) methods. The nature of these differences will highlight when each of these methods is optimal and where each method performs poorly.