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

The measurement of RNA expression in biological samples is a fruitful area of study in many areas of biology including fields as diverse as cancer research and evolutionary biology. NGS-based methods are increasing in use but the errors associated with NGS-based measurements are not well understood. All existing NGS-based measurement methods are multi-step and it is unclear how errors propagate through the process to affect the final measurement. Moreover, it is commonly understood that measurements taken of low-expressed sequences are not reliable, but it is unclear where to draw the cut-off.

No current methods exist to directly compare the accuracy and precision of NGS based RNA measurement systems or to evaluate the sources error. ***We hypothesize that relative sensitivity can be adapted to provide a novel method for evaluating the error associated with next-generation sequencing technologies.*** Relative sensitivity was originally developed by John Mandel in 1985 for evaluation of measurements in analytical chemistry. The method has been largely ignored in the evaluation of modern methods but has several useful properties. For one, relative sensitivity is invariant to differences in scale and monotone transformations. This makes it an ideal candidate for comparing measurements that utilize different scales or data transformations, such as RNA-seq. We believe the theory of relative sensitivity will provide a powerful framework for comparing measurement methodologies without the need for individual calibration curves or known analyte concentrations.

Aim 1: Develop relative rensitivity method for NGS-based count data.
We will reformulate the original relative sensitivity model to accommodate the count data associated with NGS measures. The negative binomial distribution is often used to model these data but several other formulations may provide better fit, such as a Poisson or zero-inflated Poisson. Since the true distribution of NGS count data is unknown we will re-formulate the relative sensitivity model under all 3 distributions. We will then assess model fit and utility using simulated RNA-seq data from all three models. This will enable us to characterize the bias and error around the relative sensitivity measure and determine which model formulation: 1) is most robust to violations of its underlying distributional assumption, 2) provides the best estimates of precision for high, medium, and low expressing probes, and 3) has the most stable estimates over the entire range of measurement.

Aim 2: Evaluate popular normalization procedures used in practice. It is well understood that NGS based measurements must be normalized prior to analysis. However, there are currently many normalization methods available and it is unclear which method is optimal. The only published comparison of normalization methods evaluated normalization based on the number of differentially expressed probes found in the final analysis. Not only was this result specific to differential expression analyses, it was also of limited utility due to the coarse nature of the outcome measure. We will use the estimated relative sensitivity measure to compare 6 popular normalization methods on both simulated data (from all three generating distributions above) and NGS data generated from technical replicates of murine brain RNA. Using relative sensitivity to compare these methods will uncover the relative ways each method propagates and mitigates measurement error and batch effects.

Aim 3: Construct Relative Limit of Detection and Relative Limit of Quantitification metrics. We will also use relative sensitivity to construct two currently undefined metrics for NGS measurement systems: the relative limit of detection (LOD) and the relative limit of quantification (LOQ). Standard estimates for the limit of detection and limit of quantification are not applicable to NGS data. We will use relative sensitivity to formulate new estimates for relative LOD and LOQ for NGS measurement data. Currently, probes which are expressed at a low level are generally considered unreliable and ignored for further study. By providing a tool which will identify these limits we will provide important information for developers to improve them. We believe the lack of published effects among low expressors in the current literature is, in part, due to the excessive measurement error in this range.

We will create an R package that provides comprehensive functions to handle NGS measurement data and implement all of the methods mentioned above. We will also create a free website using R and the shiny R package to create a user friendly and secure graphical user interface for users to easily upload data and implement our methods. We believe a website provides the most accessible form of methodological implementation to a broad audience.

Future advancements in the use of measured RNA for understanding biological systems depends on reliable measurements, particularly as the effect sizes researchers are looking for decrease as is typical in a maturing field. The necessary first step towards reliable measurements is understanding where errors arise and what factors influence them. Our proposed method will enable researchers to directly and easily compare the precision of competing NGS based measurement systems. We will provide accessible tools to evaluate the error of each step in the measurement process such as: the sample preparation method, the normalization method, and the type of sample input. We will publish a comparative analysis

of 6 commonly used normalization methods so researchers can use optimal normalization methods. Finally, our proposed method of evaluation will not require the use of samples with known analyte concentrations which is currently required and prohibitively difficult.