A Method for the Quantitative Evaluation of Normalization Methods Applied to RNA-Seq Data

Dominic D LaRoche

September 2, 2015

Contents

0.1 Hypothesis

Relative sensitivity (Mandel 1984) will provide an informative metric for evaluating the measurements of next-generation sequencing technologies without the burden of creating samples with known quantities of target sequences. The construction of individual calibration curves will also assist in identifying the limit of detectiona and the limit of quantitation for a particular technology.

Chapter 1

Introduction

Major breakthroughs in personalized medicine have been made possible through the quantification of individual transcriptomes through a next-generation sequencing technique known as RNA-Seq. RNA-Seq data provides insight into the cellular activity of an individual by quantifying the messenger (mRNA) or micro RNA (miRNA) present in the cellular matrix at the time of sampling. This information can be used to assess disease state, drug eligability, or even ...

1.1 Background

1.1.1 RNA-Seq Data

RNA-Seq data is created by isolating the RNA from the cellular matrix... Turning it into cDNA ... sequencing... and quantifying...

The recent proliferation of RNA-Seq technology has enabled widespread use of this method for both research use and clinical diagnostic procedures. However, raw RNA-Seq data is not immediately comparable accross studies or even accross samples within a study due to sample and run specific differences in library size, read-depth, and ... Due to this well-known problem it is common practice to normalize RNA-Seq data so that the data can be compared across samples and studies.

RNA-Seq data is also characterized by non-normal distribution of the data (counts) which include a large number of zeros

1.1.2 Normalization Methods

The practice of normalization generally refers to the transformation of data to make data comparable across samples and experiments. A number of popular normalization methods exist for RNA-Seq data including "Total Count" (TC) and "Reads per Kilobase-Pair per Million" (RPKPM). Recent comparisons have shown poor performance for both TC and RPKPM normalization methods (Dillies et al. 2012) so these methods are not considered further. However, a number of normalization methods remin available to researchers including: Quantile, Median, ...

Quantile Normalization

1.1.3 Upper Quartile Normalization

Median Normalization

DESeq Normalization

TMM Normalization

1.1.4 Current Evaluation of Normalization Methods

1.2 Purpose and Scope

Chapter 2

Methodology

2.1 Quantitative Evaluation of Normaliations Methods

2.1.1 Maximum-Likelihood Evaluation of Transformations

2.1.2 Non-identifiability

2.1.3 Jacobian

Chapter 3

Planned Dissertation Work

3.1 Paper 1: Theory of Quantitative Assessment of Normalization Methods

3.2 Paper 2: Simulation study and comparison with real data

3.3 Paper 3: R package NormComp

3.4 Methods