Comparability and reproducibility of biomedical data

Raphael Gottardo and Yunda Yuang

June 20, 2012

Abstract

1 Introduction

Over the past two decades, the biomedical field has been transformed by the advent of new high throughput technologies including gene expression microarrays, protein arrays, flow cytometry and next generation sequencing. Because these technologies generate large high dimensional data from a single machine run, data management and analysis have become an integral part of any scientific experiment. In addition to the experimental protocol, data analysis contributes siginificantly to the reproducibility or non-reproducibility of an experiment or publication. Unfortunately, as of today, too many published studies remain un-reproducible due to the lack of sharing of data and/or computer code or scripts that can be used to reproduce the analysis. This lack of reproducibility has even gone has far as to stop a cancer clinical trial based on gene expression signatures that could not be reproduced by independent researchers. Should have the data and computer code been made available, the results of the study could have been invalidated more rapidly, which could have saved funding and avoid giving patients false hope.

2 Reproducibility of assay and primary data

2.1 Overview of data generation process and its impact on C&R

2.2 Metrics to quantify C&R

[Figure 1 about here.]

2.3 Correcting for experimental bias

2.4 Standards and data sharing

3 Reproducibility of assay results and derived data

Here we discuss some of the tools available to researchers to perform reproducible analysis and share processed data, computer code and final results.

List of Figures

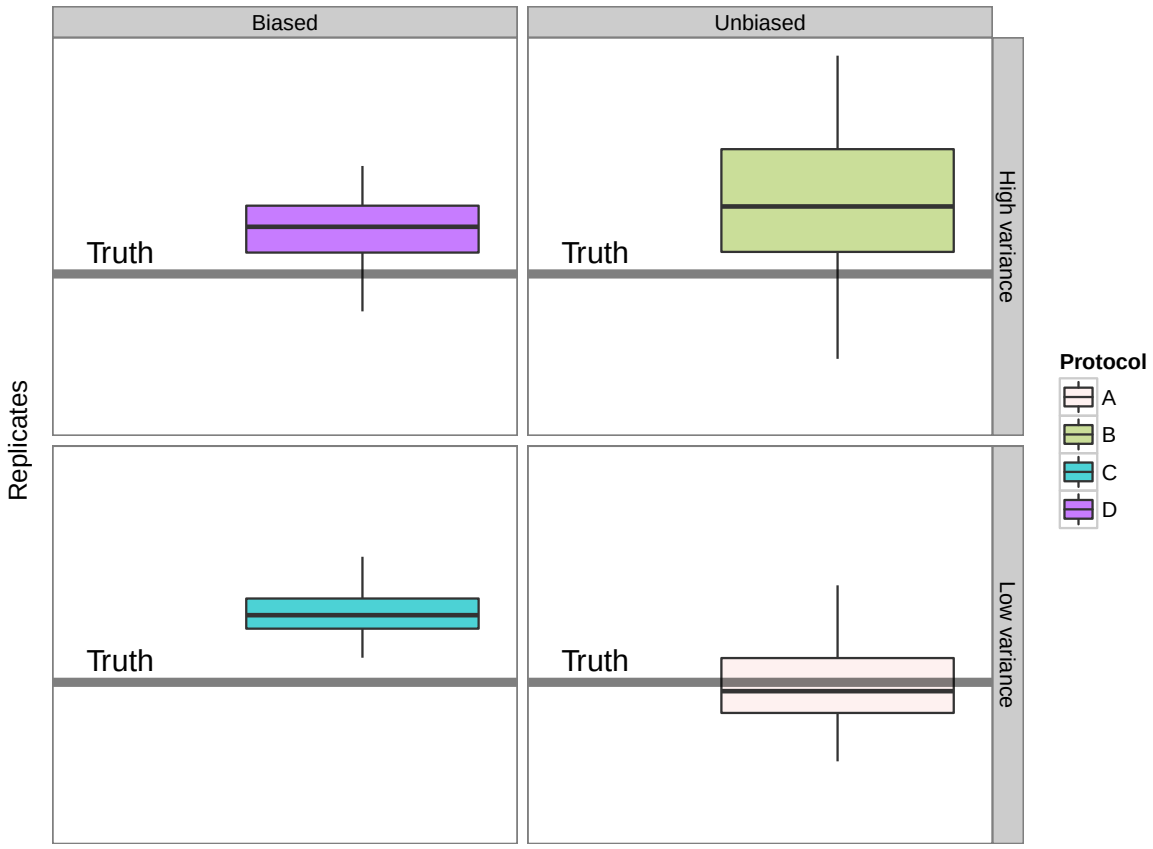

Figure 1: Variance-Bias trade off.