

# Comparability and reproducibility of biomedical data

Raphael Gottardo and Yunda Yuang

June 26, 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 significantly 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 as 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. Fortunately, over the past decade computers, software tools and online resources have drastically improved to the point that it is easier than ever to share data, code and construct fully reproducible data analysis pipelines.

In this paper we review some of the fundamental issues involved in the reproducibility and comparability of biomedical data going from assay standardization to reproducible data analysis. This paper is not meant to be an exhaustive review of all possible assays and problems but rather to select a few concrete examples and present some thoughts and solutions towards the overall C&R concept. The paper is organized as follows

## 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.

### **3.1 Tools for reproducible analyses**

### **3.2 Standards and code sharing**

### **3.3 Authoring tools**

## **4 Conclusion**

## List of Figures

1	Variance-Bias trade off. . . . .	4
---	----------------------------------	---

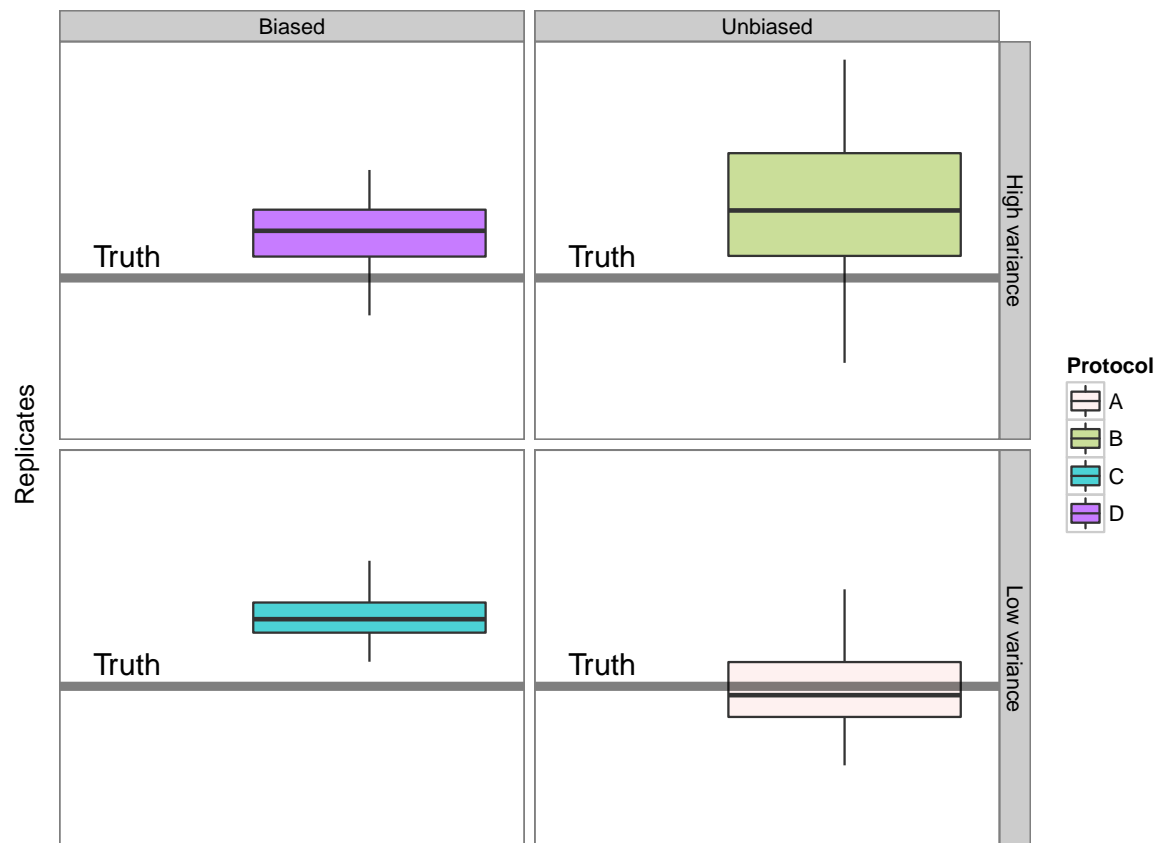


Figure 1: Variance-Bias trade off.