# Comparability and reproducibility of biomedical data

Raphael Gottardo and Yunda Yuang

June 27, 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. 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 fondamental 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

In the same fashion that experimental protocols need to be published in order for an experiment to be reproduced, computer code and data should also be published along with the results of a data analysis. In order to facilitate exchange and reproducibility, code and data should be standardized as much as possible. to embed the R code for complete data analyses in latex documents

### 3.3 Authoring tools

Several tools have been proposed to automatically incorporate reproducible data analysis pipelines or computer code into documents. An example is the GenePattern Word plugin that can be used to embed analysis pipelines in a document and rerun them on any GenePattern server from the Word application. Another example that is popular among statisticians and bioinformatics is the Sweave literate language that allows one to create dynamic reports by embeding R code in latex documents. This is also our prefered approach because it is open source and does not depend on propriety software. In addition, recent software development such as RStudio and knitr made working with Sweave even more accessible, which should reduce the learning curve for most users. In fact, this article was written using the Sweave language and processed using RStudio. Ideally, all material including the Sweave source file, computer code and data, which Gentleman refers to as a compendium, would be made available along with the final version of the manuscript and be open access, allowing anyone to reproduce the results or identify potential problems in the analysis. This openness should further improve the impact of open access papers and journals over non-open access journals by giving more credibility to the published results. Unfortunately, currently very few journals are pushing for reproducibility and even less have clear reproducibility policies. An example of a journal moving in the right direction is Biostatistics, for which one of us is associate editor. Biostatistics now has a reproducility guideline and is now working with authors towards making sure that published results are reproducible given that data and code are provided (as described in the guileline). When data and code are provided, and results are reproducible, the article is marked with an R.

### 3.4 Open science

## 4 Conclusion

# List of Figures

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

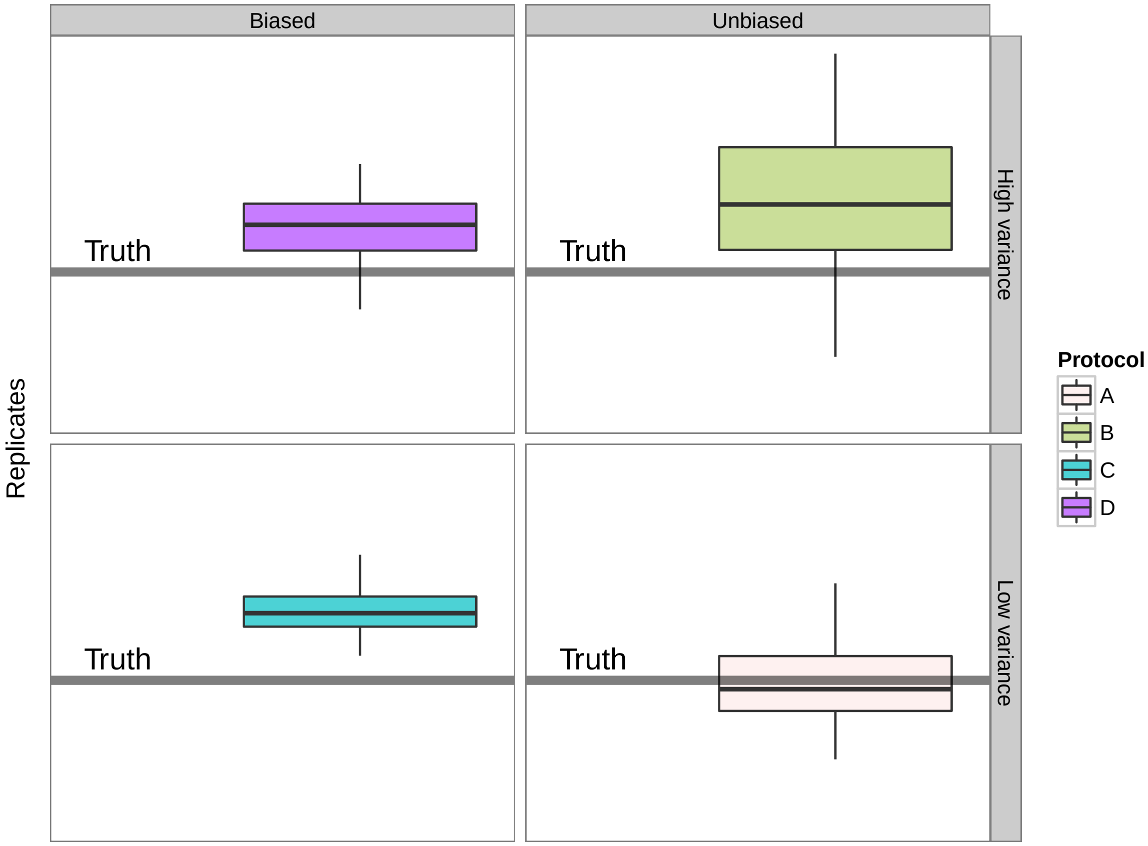

Figure 1: Variance-Bias trade off.