BRIEFINGS IN BIOINFORMATICS. VOL 14. NO 4. 391–401
Advance Access published on 27 November 2012
doi:10.1093/bib/bbs078

Comparability and reproducibility of biomedical data

Yunda Huang and Raphael Gottardo

Submitted: 10th July 2012; Received (in revised form): 18th September 2012

Abstract

With the development of novel assay technologies, biomedical experiments and analyses have gone through substantial evolution. Today, a typical experiment can simultaneously measure hundreds to thousands of individual features (e.g. genes) in dozens of biological conditions, resulting in gigabytes of data that need to be processed and analyzed. Because of the multiple steps involved in the data generation and analysis and the lack of details provided, it can be difficult for independent researchers to try to reproduce a published study. With the recent outrage following the halt of a cancer clinical trial due to the lack of reproducibility of the published study, researchers are now facing heavy pressure to ensure that their results are reproducible. Despite the global demand, too many published studies remain non-reproducible mainly due to the lack of availability of experimental protocol, data and/or computer code. Scientific discovery is an iterative process, where a published study generates new knowledge and data, resulting in new follow-up studies or clinical trials based on these results. As such, it is important for the results of a study to be quickly confirmed or discarded to avoid wasting time and money on novel projects. The availability of high-quality, reproducible data will also lead to more powerful analyses (or meta-analyses) where multiple data sets are combined to generate new knowledge. In this article, we review some of the recent developments regarding biomedical reproducibility and comparability and discuss some of the areas where the overall field could be improved.

Keywords: *Analysis pipeline; accuracy; open science; precision; protocol; standardization*

INTRODUCTION

Over the past two decades, the biomedical field has been transformed by the advent of new high-throughput technologies such as gene expression microarrays, protein arrays, flow cytometry and next-generation sequencing. Experiments and protocols have become increasingly complex, involving the use of instruments that can be very sensitive to specific settings. For example, small changes in the photomultiplier tube voltage of a flow cytometer or a microarray scanner could drastically change the output of an experiment [1]. It is thus crucial that protocols be well described, standardized and shared in order for an experiment to be reproducible and comparable within and between laboratories.

Furthermore, these novel biomedical technologies generate large high-dimensional data sets from individual experiments. The growth of such data has highlighted the importance of implementing data management and analysis plans as an integral part of experimental design. In consequence, data analysis procedures contribute significantly to the reproducibility or non-reproducibility of an experiment or publication. Unfortunately, as of today, too many published studies remain irreproducible due to the lack of sharing of data, computer code or software required to reproduce the study results. This lack of reproducibility has had significant impact, leading to the halt of a cancer clinical trial when key gene expression signatures used for decision making were

Corresponding author. Raphael Gottardo, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Mailstop M2-C200, Seattle, WA 98109-1024, USA. Tel.: 206-667-4076; Fax: 206-667-4378; E-mail: rgottard@fhcrc.org

Yunda Huang specializes in the design and analysis of pre-clinical and clinical vaccine studies. She is currently Senior Staff Scientist at the Fred Hutchinson Cancer Research Center.

Raphael Gottardo specializes in the development of statistical methods and software tools for the analysis of high-throughput and high-dimensional biological assays. He is currently an associate member at the Fred Hutchinson Cancer Research Center and an affiliate associate professor at the University of Washingon.

found to be caused by analysis errors and could not be independently reproduced by researchers [2]. Had the data and computer code been made available, the results of the study could have been invalidated more rapidly, which could have saved funding, avoided giving patients false hope and most importantly ensured patients received effective treatment [3]. 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 article, we review some of the fundamental issues involved in the comparability and reproducibility (C&R) of biomedical data going from assay standardization to reproducible data analysis. Our intent is not to exhaustively review all possible problems with all existing assays, but rather to select a few concrete examples based on our own experience and present some thoughts and solutions toward the overall concept of C&R. This article is divided into two main sections, one related to the experiment reproducibility and one to the analysis reproducibility, though the two topics significantly overlap.

REPRODUCIBILITY OF ASSAY AND PRIMARY DATA

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

We examine a prototypical biomedical data generation process to illustrate factors that may negatively impact the C&R of the data throughout different stages of the process. As shown in Figure 1, a data generation process can be roughly broken down into three core stages (Steps 1–3) of information transformation from signals contained in biological samples to numeric values captured in data sets for analysis. In Step 1, biological samples are measured and raw instrument data are generated. There are

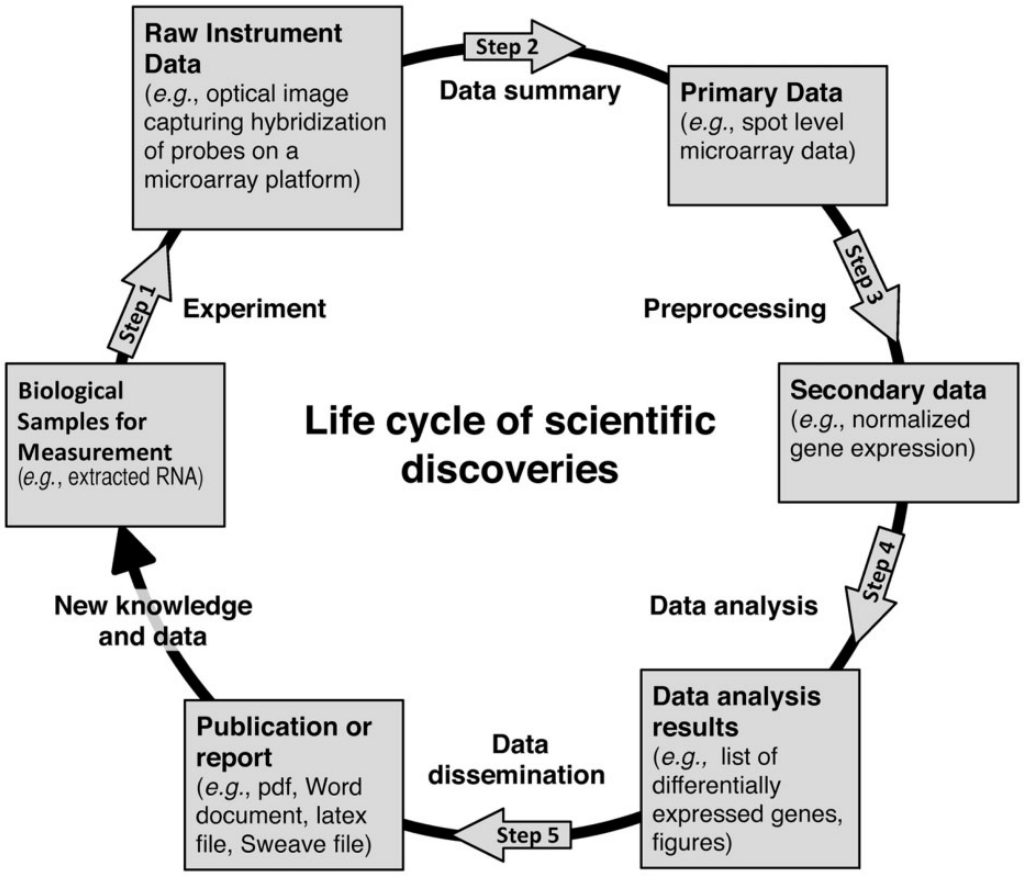

Figure 1: Life cycle of scientific discoveries. The overall cycle is broken down into five different steps. After completion of all steps according to the reproducible guidelines (Table 1), the results would rapidly lead to confirmed (or discarded) discoveries. The confirmed discoveries would then be translated into new knowledge and data supporting novel studies.

several factors that may influence the C&R of data at this stage. These include some obvious factors such as the specific type of technologies (e.g. hybridization-based or sequence-based gene expression) [4–7] or platforms (e.g. Affymetrix, Illumina or Operon) [8–12], the Standard Operating Procedures (SOPs) for biological sample preparation, experimental design, experiment layout and measurement [13, 14], as well as other conditions that are often not specified in the experiment protocol. For example, the level of experience or expertise of the technicians performing the experiment [15, 16], or the origin of the reagents (e.g. batch effects [17, 18]) are also possible sources for differences between independent experimental results. Therefore, in Step 1, to increase the C&R of data, all these factors should be thought out and optimally controlled and standardized whenever possible. When factors such as technicians or reagent batches may not be standardizable across multiple studies or laboratories, a measuring system comprised of a specific platform using a specific technology should strive to minimize variations caused by these factors and increase robustness against changes in these factors. Whenever possible, the SOPs should be shared and made available to the community. Several online platforms are now available for storing and sharing such information including 'elabprotocols' (elabprotocols.com) and 'figshare' (figshare.com). In Step 2, raw information from an instrument is calibrated and quantified into numeric values. This step often involves image analyses for information alignment and/or dimension reduction. Consequently, the specific algorithms used to make such transformations, their implementation in software and the specific data storage structures, including data formats (i.e. databases or flat files) and variable naming conventions, are vital to maintaining data consistency and should be standardized and recorded to a maximal level for effective C&R of the data. We will refer to the data derived from this step as primary data versus the secondary data generated after Step 3. In some specific cases, primary data are derived directly from the instrument, but in many cases the extremely large size of the raw data (e.g. raw images) makes it prohibitive to share these and the lack of true raw data is accepted. In 'Standards and Data Sharing' section, we provide more discussions on data standards and data sharing. Finally, in Step 3, data from Step 2 are further (pre-) processed before study objective-driven analyses are conducted. This later step often involves further data alignment such as background adjustment or data aggregation such as per-biomarker summarization from multiple subset measurements. Certain quality assurance and control processing may also occur to remove unreliable data and reduce any systematic variations between data points. As in Step 2, the specification and implementation of the algorithms and the data storage structures should be tracked in the effort to maintain the C&R of the data. In 'Reproducibility of Assay Results and Derived Data' section, we will discuss some of the tools available to share Step 2 data and associated computer code for data processing and analysis.

Metrics to quantify C&R

We use accuracy and precision as two building-block metrics to illustrate the concept of C&R. While the exact definition of C&R may vary depending on the context, accuracy and precision are two well-defined statistical concepts. Specifically, accuracy indicates how close a measurement is to its true (actual) value, whereas precision indicates how close measurements are to each other. Deviation from accuracy (i.e. bias) is often introduced by systematic sources of error. For example, factors mentioned earlier such as the measuring system or a poor reagent may be a primary source of bias that cannot be removed by repeating or averaging large numbers of measurements. On the other hand, precision (i.e. variability) of data can generally be improved by increasing the number of measurements. For this reason, biological and technical replicates are recommended in an experimental design to help distinguish biological variation from technical variation. In general, there is a trade-off between accuracy and precision, in the sense that one cannot optimize both simultaneously. For example, in microarray image analysis, spots can either be summarized by the estimated foreground intensity or the background-corrected intensity (foreground minus the background). Foreground intensities are typically less variable but can exhibit higher bias compared with background-corrected intensity. In this context, many research groups have proposed pre-processing techniques that aim at finding a good compromise between the two [19]. A hypothetical example is shown in Figure 2, where comparable and reproducible data do not necessarily require unbiased measurements as long as they are 'consistently inaccurate' (Panel C). Imagine a hypothetical gene expression device that

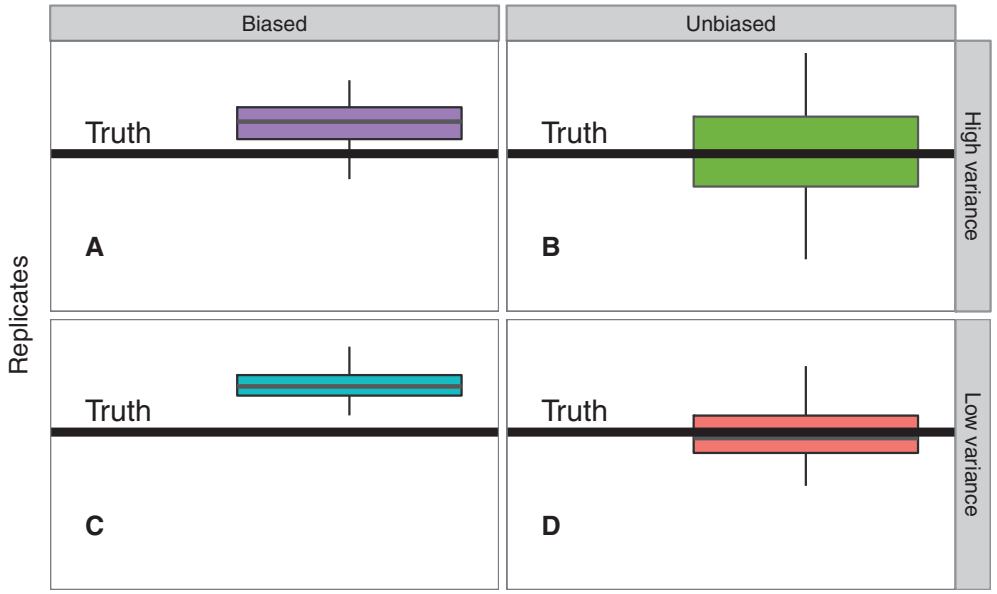

Figure 2: Precision-accuracy trade off. Four different protocols are compared. Protocol B exhibits large variance (wide box) with small bias (close to the true value on average) while protocol C has small variance but large bias. Overall, protocol D exhibits good variance-bias trade off and should be prefered.

always measures the expression of a gene as being zero. The experiment is highly reproducible but completely biased and thus useless. It is not atypical for an experimentalist to compute a coefficient of correlation between two series of experiments and to be very pleased when he/she obtains a value close to 1. Unfortunately, the large correlation could be explained by the fact that the measurements are biased and both are correlated with the same experimental artifact. So it is important that when C&R is evaluated, accuracy is also taken into consideration. Therefore, to ensure meaningful integrative analysis of biomedical data from multiple sources, although there may be issues of reliability, we encourage the inclusion of a well-established 'gold standard' of measurement whenever possible such as the inclusion of 'established' positive and negative controls that provide reasonable upper limits on the sensitivity and specificity of the experimental measurements. In this way, any signals identified from comparable and reproducible data can also be scrutinized against the gold standard for true scientific values.

Methods to correct for experimental bias

In the presence of possible experiment-specific bias, data pre-processing methods can be used to improve C&R. It is common practice to reduce non-biological sources of variation via pre-processing techniques such as background correction, batch effect removal or normalization. Many of these methods were established during the early days of microarrays at a time when experimental procedures were still being optimized and technical variability was omnipresent. Such methods include lowess normalization [20], quantile normalization [21], ComBat [22], SVA [23] and RUV-2 [24] for batch effect removal and gcRMA for removing non-specific binding of oligonucleotides [25], to cite a few. Due to the positive impact these methods have had on C&R, many other fields have adopted similar pre-processing techniques, e.g. flow cytometry [26] and next-generation sequencing [27]. Most of these methods rely on the assumption that the majority of biomarkers (genes or proteins) are not differentially expressed and the numbers of up-and down-regulated biomarkers are roughly equal across samples. Such an assumption can be reasonable when the dimension of the biomarkers collected in each sample is large but may not be satisfied in lower dimension biomedical data. In the latter case, internal or external validation data are usually used to correct for experimental bias that may be related to measurement, instrument or sampling design [28]. When there is a lack of standard for a quantity's true value [29] and validation data are infeasible to generate, calibration methods based on paired samples [30] can be adopted to adjust for experiment bias. For example, in the field of flow cytometry true gold

standards do not exist yet and it is thus difficult to evaluate C&R. The FlowCAP group (flowcap.flow-site.org) is currently working with the Human Immunology Project group [31] to derive objective criteria and gold standards that will be used to standardize and evaluate pre-processing of flow cytometry data.

Standards and data sharing

As data sets get richer with more data points, more variables and more metadata, it is important to define standards that can be used to capture and distribute all necessary information toward achieving reproducibility [32]. Several standards have been proposed for biomedical data to achieve these goals including MIAME for gene expression [33], MINSEQE for sequencing experiment [34], MIATA for T cell assays [35] or MiFlowCyt for flow cytometry [36]. In addition to assay protocol information, primary and secondary data, it is important that any pre-processing done to the data be fully described (e.g. normalization for microarrays). Unfortunately, too many assays are still lacking data standards (e.g. bead array multiplex assays) or if data standards are available, manufacturers and/or software companies have been slow at adopting them. For example, despite the availability of data standards for defining preprocessing for flow cytometry, no analysis software has yet fully adopted this format and it is very difficult to share reproducible analyses across software platforms. We, the flow informatics community, basically had to reverse engineer commercial software file formats and write custom open-source software that can read these [37].

Funding agencies have been very supportive to the creation and adoption of standards for biomedical data, by funding many of the standards that are existing. For example, as part of the Human Immunology Project Consortium (HIPC), a project funded by the NIH, we and other bioinformaticians are currently working toward the definition of novel standards for immunological data. Similarly, the Collaboration for AIDS Vaccine Discovery (CAVD), funded by the Bill and Melinda Gates Foundation (BMGF), has set up an immune monitoring consortium to establish validated T-cell and antibody immunological assays across a network of Good Clinical Laboratory Practices-certified laboratories that could monitor the anticipated pipeline of HIV vaccine trials emanating from the field. Once data and data formats have been standardized, it is important to make these data publicly available for the benefit of science, and to this extent, funding agencies have an important role to play. Most funding agencies including the National Science Foundation and the NIH clearly encourage investigators to share data and/or have defined policies to this end. Similarly, charitable organizations such as the BMGF and the Wellcome Trust are also actively working with grantees to maximize the amount of data available to the research community. Example projects that have good data sharing policies and have setup databases for sharing data, that we are personally involved in, are the HIV Vaccine Trials Network (HVTN), HIPC and the CAVD. In addition to helping retrieve data more efficiently (e.g. via queries), databases can help minimize human errors in data manipulation by ensuring that raw and processed data along with metadata are automatically uploaded with minimal manual intervention. Databases can also help maintain data consistency by checking that some standards are followed or by doing basic data quality checks. For example, the Immunological Portal database (ImmPort.org) provides data templates that help investigators upload their data in a standardized format. It is thus a good idea to use specialized databases whenever possible to store and share data. Despite this global effort, many policies are still either too vague or not properly enforced and data are treated as the private property of investigators who aim to maximize their publication record at the expense of the widest possible use of the data. This situation threatens to limit both the progress of the related research and its application for public health benefit. We feel that it is important for funding agencies to set stricter and clearer data sharing policies, particularly for sensitive data (e.g. individual genomes and clinical data) where policies are often vague or industrial partnerships make the creation of such policies very difficult. In these cases, despite their sensitive nature, these data could and should be shared as long as they are properly de-identified to protect the patients identity under the Health Insurance Portability and Accountability Act.

Once data and all necessary information are made available, these data need to be appropriately cited when the study and its results are published. To this end, it is crucial that journals set data sharing policies or guidelines and that authors do follow these guidelines. Unfortunately, as mentioned in a recent study [38], too few journals have clear policies for data deposition and even fewer make it mandatory for publication. That study found that even when data

deposition is a requirement, the majority of authors did not fully follow the instructions. For example, it is common for researchers to share processed data only, which makes it nearly impossible to reproduce the results or use different analysis tools that require primary data. For example, in the field of genomics, many researchers share processed sequence file formats (e.g. wiggle files), which prevents anyone from analyzing the data with an algorithm that requires primary data (e.g. raw or aligned reads).

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 as detailed in the following subsections and summarized in Table 2. Analysis of data issued from high-throughput experiments can be extremely complex, involving multiple steps from data formatting and pre-processing to statistical inference. Thus, it is important that all steps be recorded for full reproducibility as shown in Figure 1 and Table 1 (Steps 2–5). This can be difficult to do with a point-and-click software interface, where there is no easy way to save intermediate results. This is not to mention the fact that the 'manual' analysis of a high-throughput data set typically requires the use of multiple software tools and is very time consuming. In addition, it is not clear how robust the conclusions of a study are to small perturbations in any of these analysis steps. As such, it might be a good idea to be able to quickly redo an analysis after tuning some parameters to optimize the analysis; something that is not practical within a point-and-click environment.

Tools for reproducible analyses

In recent years, several open-source, community-based projects have emerged that enable researchers to construct and share complete and fully reproducible data analysis pipelines. The Bioconductor project [39], based on the R statistical language [40], provide >500 software packages for the analysis of a wide range of biomedical data, from gene expression microarrays to flow cytometry and next-generation sequencing. These packages can be combined via scripts written in the R language to form complex data analysis pipelines, connect to data repositories and generate high-quality graphics. The resulting R scripts can then be used to record and later reproduce the analysis (along with all input parameters). Because all steps of the analysis are automated when the script is executed, it is easy to assess the robustness of the results when tuning some parameters. Other similar projects with perhaps more focused capabilities include BioPython [41] and BioPerl [42] that are based on the Python and

Table 1: Checklist for a comparable and reproducible experiment following stages in the life cycle of scientific discoveries as shown in Figure 1

Scientific discovery stage	Recommendations	Check
Step 1: Biological samples for measurement	Store and share source of samples and/or samples if possible	☐
	Store and share extra samples for reproducibility (when possible/applicable) and future studies	☐
Step 2: Raw instrument data	Standardized experimental protocol	☐
	Store and share measuring system (technology and platform)	☐
	Store and share Standard Operating Procedure (SOP)	☐
	Store and share experiment conditions not specified in SOP (e.g. techinician and time)	☐
Step 3: Primary data	Perform quality control	☐
	Store and share primary data and metadata	☐
	Store and share code and softwarefor algorithms used during summary (e.g. image analysis)	☐
	Use open-source software and avoid point-and-click analysis interfaces	☐
	Use data standards and databases	☐
Step 4: Data analysis results	Store and share analysis results and derived data	☐
	Store and share code and software (with versions)	☐
	Use open-source software and repository for sharing code and data	☐
	Validate results using independent data or experiment(s) (when possible)	☐
Step 5: Publication or report	Publish results with link to code, data and software	☐
	Use dynamic reporting when possible (e.g. Sweave)	☐
	Publish in open access journals	☐

Table 2: List of tools and resources for reproducible biomedical data analysis mentioned in this review

Name	Description/usage	URL
Online protocol storing and sharing		
elabprotocols	Web-based Laboratory Protocol & SOP Management	http://www.elabprotocols.com
figshare	Web-based tool for storing and sharing all sorts of research output	http://www.figshare.com
Databases and data management tools		
LabKey Server	Biomedical research data management with powerful programming interfaces for analysis	http://www.labkey.com/
ImmPort	The Immunology Database and Analysis Portal	http://www.immport.org
Analysis tools		
Bioconductor	Collection of R packages for high-throughput biological data analysis	http://www.bioconductor.org/
Biopython	Python tools for computational molecular biology	http://www.biopython.org
BioPerl	Perl tools for bioinformatics, genomics and life science research	http://www.bioperl.org
Analysis platforms with graphical user interface		
RStudio	Integrated development environment (IDE) for R	http://www.rstudio.org/
GenePattern	Genomic analysis platform with web-based interface	http://www.broadinstitute.org/cancer/software/genepattern/
GenomeSpace	Genomic analysis platform linked with multiple tools including GenePattern, Galaxy and Cytospace	http://www.genomespace.org/
Code sharing and versioning tools		
GitHub	Web-based tool for software development and collaboration based on the Git version control system	http://www.github.com/
Authoring tools		
GenePattern Word Plugin	Microsoft Word add-in for the GenePattern Reproducible Research Document	http://www.broadinstitute.org/cancer/software/genepattern/
Sweave	Integration of R code into LaTeX documents	http://www.statistik.lmu.de/~leisch/Sweave/
knitr	Elegant, flexible and fast dynamic report generation with R. knitr is integrated in RStudio for ease of use.	http://yihui.name/knitr/

Perl languages, respectively (to our knowledge, neither BioPython nor Perl have tools for the analysis of flow cytometry data).

Even though several graphical user interfaces (e.g. RStudio for R) are available for writing computer scripts based on R/Bioconductor (or BioPerl, BioPython), the learning curve can still be steep for novice users. More user-friendly-based tools are now available to construct reproducible data analysis pipelines using combinations of available modules that are for the most part wrappers of packages written in R, Perl or Python (or some other language). For example, a popular platform for gene expression analysis, GenePattern, versions every pipeline and its methods, ensuring that each version of a pipeline (and its results) remains static [43]. A more recent project, GenomeSpace (genomespace.org), funded by the National Human Genome Research Institute, can now combine GenePattern with other popular Bioinformatics tools including Galaxy, Cytoscape and the UCSC genome browser. As such, users can perform all of their analysis using a single platform. In the clinical and immunological field, LabKey Server is a popular web-based tool for storing immunological data (via a database) and building complex analysis pipelines that can be shared with other users [44]. LabKey Server also versions every pipeline for full reproducibility. LabKey Server is currently being used by large research networks including the CAVD, the HVTN and the Immune Tolerance Network, to name a few.

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, software and data should also be published along with the results of a data analysis. Ideally, software would be open source and computer code would be well packaged and standardized to facilitate exchange and usability. Both Bioconductor and GenePattern, mentioned earlier, provide facilities for users to package and share code with other users. Bioconductor is based on the R packaging system, which is highly standardized and has been a driving force behind the wide adoption of both R and Bioconductor.

Bioconductor goes even further by: (i) ensuring that all submitted packages are peer-reviewed and (ii) providing version control repositories and build systems where source code is maintained, versioned and binaries automatically built for all computer operating systems. Among other things, the peer-review process ensures that the package follows some basic guidelines, are well documented, work as advertised and are useful to the community. The open-source and versioning system provides full access to algorithms and their implementation, which are crucial to obtain full reproducibility. For users who want to version and share software code outside of the Bioconductor (or similar) project, there exist many, free web-based hosting services to store, version and share code (and even data). One of our favorite platforms is GitHub, which the company markets as 'Social Coding for all'. GitHub makes it easy for anyone to store and version control computer code, packages, documents, webpages and even wikis to document their code. The social aspect of GitHub makes it easy for users to work in teams on a common project, software or manuscript. GitHub is free for all open-source projects.

Unfortunately, very few journals have code/software sharing policies and even fewer have requirements that the code/software be open access. For example, *BMC Bioinformatics* only has policies for software articles and even for these the source code is not required, only an executable. *PLoS One* requires authors of manuscripts in which software is the central part of the paper to release software and make code open source for submission. Although this policy is clearer, it is still up to the editor/reviewers to decide whether software was a central part of the paper. In a day and age where most experiments generate large amount of data, software is always going to play a central role, so why not make this policy universal for all submissions involving data analysis? Fortunately, based on our own experience, we feel that reviewers are pushing in the right direction by asking that code be open source and released along with the paper. So even if journals have no clear policies yet, we, the community, can enforce that code be released every time we review a paper.

Validation and robustness of results

In addition to ensuring reproducility of assay data and results, it is always a good idea to try to validate the results of a study using an idependent platform or data set. This is particularly relevant for studies involving large data sets that can generate long lists of novel findings such as a list of differentially expressed genes from a microarray experiment or a list of transcription factor binding sites from a ChIP-Seq (chromatin immunoprecipitation followed by sequencing) experiment. In the context of gene expression or ChIP-Seq, quantitative polymerase chain reaction (qPCR) can be used to validate some of the genes or sites [45, 46]. Note that such experimental assays (including qPCR) are also subject to variation, which can affect the validation [45]. If direct experimental validation is not feasible, computational validation can be used instead. For example, the list of differentially expressed genes (or biomarkers) can be tested using an independent data set that was generated by a different group. In the context of ChIP-Seq *de novo* motif finding tools have been used to validate binding sites that contain the expected motifs [47].

The lack of validation partially explains why very few published biomakers have clinical utilities [48]. In addition, when it gets to statistical inferences, robustness in model building and stability in feature selection due to sampling variations may also contribute greatly to the reproducibilty of analysis results. Several schools of intensive research have been dedicated to this area lately. For example, data mining or high-dimensional data analyses methods that incorporate resampling techniques, e.g. bagging [49] or boosting [50], often provide more stable and hence more reproducible results [51]. Similarly, predictions based on consensus of multiple analysis results are generally more robust and perform better than any single method [52].

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 [53]. Another example that is popular among statisticians and bioinformatics is the Sweave literate language [54] that allows one to create dynamic reports by embedding R code in latex documents. This is our preferred approach because it is open source and does not depend on proprietary software. As an example, every Bioconductor package is required to have fully reproducible documentation (called a

vignette) written in the Sweave language. Recent software development tools such as RStudio (rstudio.org) and knitr (yihui.name/knitr) have 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 and the source file (along with all versions of it) is available from GitHub (http://github.com/raphg/BiB-review-CR). Ideally, all material including the Sweave source file, computer code and data, which Gentleman and Temple refers to as a 'compendium' [55], 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. An obvious option would be to package code, data and the Sweave source file into an R package for ease of distribution as is commonly done for Bioconductor data packages. Anyone could directly install this package in R and have access to all necessary materials. Journals that promote this openness should further improve their impact versus non-open journals by giving more credibility to the published results, in the same fashion that open access journals typically have greater impact factors [56]. Unfortunately, currently very few journals are pushing for full reproducibility and even less have clear reproducibility policies. An example of a journal moving in the right direction is Biostatistics. Biostatistics now has a reproducibility guideline and is now working with authors toward making sure that published results are reproducible given that data and code are provided [57]. When data and code are provided and results can be reproduced by the associate editor, the article is marked with an R for reproducible.

CONCLUSION

We have reviewed some of the key steps involved in the C&R of biomedical data going from protocols to code and data sharing. For ease of reference, Tables 1 and 2 summarize some of the ideas discussed including available resources and a checklist for a comparable and reproducible scientific discovery. Even though experiments, protocols and data analyses have become more complex than ever before, tools and methods for C&R have also significantly improved. Unfortunately, we are still far from the ideal situation where every study can be reproduced and relevant data be compared and pooled across laboratories or institutions. Besides experiment and protocol consistency, there is still a lot of work to be done in terms of data and analysis standardization that would not only improve reproducibility but also facilitate data exchange and meta analyses. Perhaps one way to achieve this is for experimental and computational groups to work together when developing novel assays, standards and analysis tools. This is something that is integral to the CAVD and HIPC projects mentioned previously. For example, both the CAVD and HIPC have bioinformatics and biostatistics and assays subcommittees that work together to optimize and standardize novel assays and analysis tools.

In terms of data, code and software sharing, we cannot yet rely on goodwill and self discipline when it comes to sharing publication material and making studies fully reproducible. As such, we feel that today, the most important step forward toward improving C&R is for funding agencies, publishers and researchers to work together by setting very strict reproducibility guidelines and policies. Such policies could potentially save a great deal of money and resources by making sure that scientific errors can quickly be discovered and corrected instead of giving birth to new scientific projects and clinical trials based on erroneous results. Of course, no one should be afraid of making their publication material available because someone might identify a flaw in the study. As Alexander Pope said, 'To err is human, to forgive is divine'; we all learn by our mistakes and this is the only way science can move forward.

Key Points

- Today, a typical experiment can simultaneously measure hundreds to thousands of individual features (e.g. genes) in dozens of biological conditions, resulting in gigabytes of data that need to be processed, analyzed and potentially reproduced.
- Multiple ongoing open-source, community-based projects have emerged that enable researchers to share study protocols, experiment constructs, resulting data sets, as well as complete and fully reproducible data analysis pipelines.
- Experimental and computational groups need to work together when developing novel assays, standards and analysis tools ensuring that all steps leading to the results of a study are optimized and reproducible.
- The availability of open-acess, high-quality and reproducible data, will also lead to more powerful analyses (or meta-analyses) where multiple data sets are combined to generate new knowledge.
- Funding agencies, publishers and researchers need to set strict C&R policies that would allow rapid revealation and correction of scientific errors instead of giving birth to new scientific projects and clinical trials based on erroneous results.

FUNDING

This work was supported by Bill and Melinda Gates Foundation [OPP1032317] and National Institutes of Health [U01 AI068635-01 and U19 AI089986-01].

References

1. Lyng H, Badiee A, Svendsrud DH, *et al*. Profound influence of microarray scanner characteristics on gene expression ratios: analysis and procedure for correction. *BMC Genomics* 2004;**5**:10.
2. Hutson S. Data handling errors spur debate over clinical trial. *Nat Med* 2010;**16**:618.
3. Baggerly KA, Coombes KR. What information should be required to support clinical 'omics' publications? *Clin Chem* 2011;**57**:688–90.
4. Yauk CL, Berndt ML, Williams A, *et al*. Comprehensive comparison of six microarray technologies. *Nucleic Acids Res* 2004;**32**:e124.
5. Liu F, Jenssen T-K, Trimarchi J, *et al*. Comparison of hybridization-based and sequencing-based gene expression technologies on biological replicates. *BMC Genomics* 2007; **8**:153.
6. Kuo WP, Liu F, Trimarchi J, *et al*. A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nat Biotechnol* 2006; **24**:832–40.
7. Git A, Dvinge H, Salmon-Divon M, *et al*. Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *RNA* 2010;**16**:991–1006.
8. Larkin JE, Frank BC, Gavras H, *et al*. Independence and reproducibility across microarray platforms. *Nat Methods* 2005;**2**:337–44.
9. Baumbusch LO, Aarøe J, Johansen FE, *et al*. Comparison of the Agilent, ROMA/NimbleGen and Illumina platforms for classification of copy number alterations in human breast tumors. *BMC Genomics* 2008;**9**:379.
10. Wang B, Howel P, Bruheim S, *et al*. Systematic evaluation of three microRNA profiling platforms: microarray, beads array, and quantitative real-time PCR array. *PloS One* 2011; **6**:e17167.
11. Liu F, Kuo WP, Jenssen T-K, *et al*. Performance comparison of multiple microarray platforms for gene expression profiling. *Methods Mol Biol* 2012;**802**:141–55.
12. Chang JW-C, Wei N-C, Su H-J, *et al*. Comparison of genomic signatures of non-small cell lung cancer recurrence between two microarray platforms. *Anticancer Res* 2012;**32**: 1259–65.
13. Al-Mulla F, Al-Tamimi R, Bitar MS. Comparison of two probe preparation methods using long oligonucleotide microarrays. *BioTechniques* 2004;**37**:827–33.
14. Ach RA, Floore A, Curry B, *et al*. Robust interlaboratory reproducibility of a gene expression signature measurement consistent with the needs of a new generation of diagnostic tools. *BMC Genomics* 2007;**8**:148.
15. Duewer DL, JWRLSM. *Learning from microarray interlaboratory studies: measures of precision for gene expression*. BMC genomics 2009.
16. Todd CA, Greene KM, Yu X, *et al*. Development and implementation of an international proficiency testing program for a neutralizing antibody assay for HIV-1 in TZM-bl cells. *J Immunol Methods* 2012;**375**:57–67.
17. Scherer A (ed). *Batch Effects and Noise in Micorarray Experiments: Sources and Solutions*. Chichester, UK: John Wiley and Sons, 2009.
18. Leek JT, Scharpf RB, Bravo HC, *et al*. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 2010;**11**:733–9.
19. Gottardo R, Raftery AE, Yeung KY, *et al*. Quality control and robust estimation for cDNA microarrays with replicates. *J Am Stat Assoc* 2006;**101**:30–40.
20. Dudoit S, Yang YH, Callow MJ, *et al*. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat Sin* 2002;**12**:111–40.
21. Bolstad BM, Irizarry RA, Astrand M, *et al*. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003; **19**:185–93.
22. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;**8**:118–27.
23. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* 2007;**3**:1724–35.
24. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* 2012;**13**:539–52.
25. Wu Z, Irizarry RA. Preprocessing of oligonucleotide array data. *Nat Biotechnol* 2004;**22**:656–8.
26. Hahne F, Khodabakhshi AH, Bashashati A, *et al*. Per-channel basis normalization methods for flow cytometry data. *Cytometry A* 2010;**77**:121–31.
27. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 2010;**11**:R25.
28. Buonaccorsi JP. *Measurement error: Models, Methods, and Applications*. Chapman & Hall/CRC, 2010.
29. Maecker HT, Rinfret A, D'Souza P, *et al*. Standardization of cytokine flow cytometry assays. *BMC Immunol* 2005;**6**:13.
30. Huang Y, Moodie Z, Li S, *et al*. Comparing and combining data across multiple sources via integration of paired-sample data to correct for measurement error. *Stat Med*; doi:10.1002/sim.5446 (Advance Access publication 05 July 2012).
31. Maecker HT, McCoy JP, Nussenblatt R. Standardizing immunophenotyping for the Human Immunology Project. *Nat Rev Immunol* 2012;**12**:191–200.
32. Quackenbush J. Data standards for 'omic' science. *Nat Biotechnol* 2004;**22**:613–4.
33. Brazma A, Hingamp P, Quackenbush J, *et al*. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 2001;**29**: 365–71.
34. Society, TFGD. *Minimum Information about a high-throughput SeQuencing Experiment—MINSEQE (Draft Proposal)*. mged.org.
35. Britten CM, Janetzki S, Butterfield LH, *et al*. T Cell Assays and MIATA: The essential minimum for maximum impact. *Immunity* 2012;**37**:1–2.

36. Minimum information about a flow cytometry experiment (MIFlowCyt) checklist (Numbered in accordance with MIFlowCyt 1.0 document). *CytometryA* 2010, Vol. 77, 813.
37. Finak G, Jiang W, Pardo J, *et al*. QUAliFiER: an automated pipeline for quality assessment of gated flow cytometry data. *BMC Bioinformatics* 2012;**13**:252.
38. Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, *et al*. Public availability of published research data in high-impact journals. *PloS One* 2011;**6**:e24357.
39. Gentleman RC, Carey VJ, Bates DM, *et al*. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 2004;**5**:R80.
40. Ihaka R, Gentleman R. R: a language for data analysis and graphics. *J Comput Graph Stat* 1996;**5**:299.
41. Cock PJA, Antao T, Chang JT, *et al*. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009;**25**:1422–3.
42. Stajich JE, Block D, Boulez K, *et al*. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* 2002;**12**: 1611–8.
43. Reich M, Liefeld T, Gould J, *et al*. GenePattern 2.0. *Nat Genet* 2006;**38**:500–1.
44. Nelson EK, Piehler B, Eckels J, *et al*. LabKey Server: an open source platform for scientific data integration, analysis and collaboration. *BMC Bioinformatics* 2011;**12**:71.
45. Morey JS, Ryan JC, Van Dolah FM. Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. *Biol Proced Online* 2006;**8**: 175–93.
46. Johnson WE, Li W, Meyer CA, *et al*. Model-based analysis of tiling-arrays for ChIP-chip. *Proc Natl Acad Sci USA* 2006; **103**:12457–62.
47. Zhang X, Robertson G, Krzywinski M, *et al*. PICS: probabilistic inference for ChIP-seq. *Biometrics* 2011;**67**:151–63.
48. Diamandis EP. Cancer biomarkers: can we turn recent failures into success? *J Natl Cancer Inst* 2010;**102**:1462–7.
49. Breiman L. Bagging predictors. *Mach Learn* 1996;**24**:123–40.
50. Freund Y, Schapire R. A desicion-theoretic generalization of on-line learning and an application to boosting. *Comput Learn Theory* 1995.
51. Dudoit S. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics* 2003;**19**:1090–9.
52. Marbach D, Costello JC, Küffner R, *et al*. Wisdom of crowds for robust gene network inference. *Nat Methods* 2012;**9**:796–804.
53. Mesirov J. Computer science. Accessible reproducible research. *Science* 2010;**327**:415–6.
54. Leisch F. Sweave, part I: Mixing R and LaTeX. *R News* 2002;**2**:28–31.
55. Gentleman R, Temple Lang D. Statistical analyses and reproducible research. *J Comput Graph Stat* 2007;**16**:1–23.
56. Eysenbach G. Citation advantage of open access articles. *PLoS Biol* 2006;**4**:e157.
57. Peng RD. Reproducible research and Biostatistics. *Biostatistics* 2009;**10**:405–8.