CORE Issues in Professional and Research Ethics, Volume 1, Paper 6, 2012
ISSN: 2160-8784
http://nationalethicscenter.org/content/article/175

Reproducible Research

Paul A. Thompson[1] and Andrew Burnett[2]

Abstract

The concept of responsible research is a relatively new one in science. It concerns the manner in which research is done, and in which the analysis of research findings are conducted. Research involves both ethical and methodological components. Reproducible research involves methods and approaches that enhance the integrity of research and promote consistent expectations within the scientific community. Reproducible research ensures that the analysis portion of a research effort can be redone with equivalent results. A series of recent problems in published and funded research has led journals, methodologically knowledgeable scientists, and federal funding agency personnel towards proposals to increase the reproducibility of the research enterprise. Thus, research methods that can be traced from data to conclusion, whether those conclusions are defined by graphical methods, tables of summarized information, statistical evaluations of the information, or use of the information in planning for subsequent studies, are increasingly being emphasized. Research reproducibility requirements include availability of the full analysis dataset, the analysis program, and all tables, figures, and incidental computations for the paper or document. Reproducibility also involves the use of clearly defined approaches for the inclusion of information in papers. The methods for reproducible research involved in several commonly used statistical analysis programs are defined and described.

[1] Director, Methodology and Data Analysis Center, Sanford Research/University of South Dakota, Sanford Health, Sioux Falls, SD
[2] Clinical Assistant Professor, Neuroscences, University of South Dakota School of Medicine, Sioux Falls, SD

Introduction

Recent discussions in the computation sciences, biostatistics and general methods, and scientific journalism have included the concept of *reproducible research*. This concept can be stated simply:

> Reproducibility involves methods to ensure that independent scientists can reproduce published results by using the same procedures and data as the original investigators. It also requires that the primary investigators share their data and methodological details. These include, at a minimum, the original protocol, the dataset used for the analysis, and the computer code used to produce the results. (Laine et al., 2007).

Reproducible research (RR) is a general principle, rather than a specific set of actions or activities. It generally refers to processes and methods which ensure that the outcomes of scientific work products, whether these are computational algorithms, scientific papers, or computational simulations, can be reproduced without any discrepancy due to the attempt at reproduction by different users or at different times. The notions of RR will be first discussed by considering ethical ideas in science. After introducing the basic idea, a few cases which illustrate the problems inherent in lack of reproducibility and which have occurred recently will be discussed. Then, a historical survey of uses, conception, and development of RR methods will be presented. Finally, some opportunities for RR development will be considered.

Ethical behavior of the scientist

Good science is ethical science

Science is our name for the activity that involves finding the truth about nature, ourselves, and the universe. The scientist endeavors to learn about the area in which she is a specialist. He poses questions that have not been answered, examines the answers of others to determine if she agrees with them, and attempts to bring his expertise to solve problems and discover truth. Among other things, ethics involves finding truth in a truthful way, or doing so in a way that is itself truthful. Part of that involves keeping accurate records, performing analyses accurately and in a manner that can be redone later, and ensuring that truths found by the scientist are correct, and not merely accidents. This involves producing and maintaining records of the

process of scientific discovery, writing good descriptions of the process, and keeping data in a careful and accurate manner.

Experiments should be repeatable, but the emphasis on repeating previous work varies from discipline to discipline. Training in the physical sciences often involves repeating experiments; for example, repeating the Priestly experiment about oxygen is pretty simple. In some areas of science, particularly in the social and behavioral sciences, repeating key findings is difficult or impossible. In some cases, the data may not be able to be recreated due to changing rules about treatment of subjects. Such is the situation with the Zimbardo experiment about prisoners and guards. In other cases, data may have been obtained from a sample that would be overly expensive or quite difficult to repeat. However, it is important, and bordering on vital, that information about the methods for repeating experiments exists. Thus, in a written description of an experiment as published in a reputable journal, the *Methods* section is always included. This key section is found after the Introduction section and states, in detail, every step of physical manipulation, chemical use, sample storage, subject recruitment, and so forth of the conduct of the scientific process. Note that the opinion of the reader about minute detail may vary, as some information (freezer temperature for long term storage, the operating system of the computer used for the analysis) is sometimes omitted. The reader is required to assume that a level of reasonable sense is used by the experimenter.

Repeatability as a function of accurate description

Here is an example Methods section of a paper of which Thompson is a coauthor:

> Subjects were participants in a randomized placebo controlled trial of escitalopram ... adults 60 years or older. All subjects had a principal diagnosis of GAD ... a score of 17 or greater in the Hamilton Anxiety Rating Scale ... free of dementia or other neurodegenerative disease .... Participants were recruited in 2005 to 2007 from primary care practices ... Antidepressant or anxiolytic coprescription was not allowed, with the exception ... The study was approved by ... Subjects were randomized to escitalopram or placebo for 12 weeks using a permuted-block, 1:1 randomized list generated by a study statistician. The starting dose was 10 mg daily; after 4 weeks, for subjects who did not achieve response, the dose was increased ... Participants were seen weekly for the first 4 weeks and then every other week. Outcome assessments for this pharmacogenetic analysis were the Clinical Global Impressions Improvement Scale (CGI14) and

> the Penn State Worry Questionnaire (PSWQ). The main outcome was the CGI ... (Lenze et al., 2010, p. 2)

This description indicates who was selected for the trial in terms of past history of mental illness, the population from which they were selected, the manner in which they were treated, the length of time required for treatment, the manner in which the participants were treated, and the types of outcome variables used to perform the comparison of groups. If another trial were to be set up in order to reproduce the results, this description would be carefully consulted. The Methods section provides a key into the scientific method and to the actual conduct of the research being reported.

A Methods section from another article:

> A radiation-induced fibrosarcoma, RIF-1, originally obtained from ... The tumours were implanted in the flanks of female C3H/HeJ mice ... Tumour volume was estimated by calculating the volume of an ellipsoid from ..... Local hyperthermia was administered by a NMR compatible inductive heating system developed in the laboratory and described in detail in a companion report. Briefly, a 4MHz radiofrequency (RF) field was imposed over the tumour volume by a surrounding ... far the temperature of the tumour was from the desired target temperature. Control of core body temperature was ... Target temperatures ... were selected for hyperthermia treatments. A heating duration of 30 min ... The timing origin of this 30 min 'heating period' was taken as the point ... Sham experiments were also performed with 31P NMR assessment of hyperthermia ... Those tumours whose heating approach an idealized profile ... were used for correlation/regression analyses. ... All NMR measurements were performed ... Mice were anaesthetized ... An RF transmitter/receiver coil ... Magnetic field homogeneity in all experiments was optimized ... Thus, a benzene solution of ... was used as an external chemical shift ... (Bezabeh et al, 2004, p. 338)

This Methods section presents, in detail, the methods by which the experiment was conducted. The source of the experimental material is given. The treatment of research subjects (mice) is stated. Methods for measurement, treatment during the experiment, and radiation administration are described in sufficient detail for reproduction by another experimenter. The Methods section also includes some discussion of deviation from the main protocol. After a careful reading of the Methods section, knowledgeable reader should be able to provide a plan to perform an equivalent experiment.

The chain of research evidence

The process of research involves posing a scientific question and then answering that question in a clear manner, in which the competing answers (threats to inference) have been controlled or ruled out to the best of the researcher's ability. In posing and answering the research question, the concept of *chain of evidence* is key. The chain of evidence is a carefully constructed series of actions ensuring that the information gathered on the laboratory bench is untainted by deletion of data, insertion of inappropriate data, or modification of data.

This concept is a familiar idea to the viewers of crime dramas. At the scene of the crime, the crime investigators wear gloves to keep their own biological traces from affecting the material found at the scene. Specimens found at the scene are bagged in clean containers which are numbered and entered into a log of information. Pictures are taken of ephemera (e.g., tire tracks, shoe marks, marks on walls, positions of items) which cannot be removed from the scene without serious disruption. In many cases, the person who gathers the evidence works with another person, who provides a second, separate indication of what was found at the location. After returning from the scene to the location in which the evidence is to be processed, a second set of controls is instituted. Evidence is kept in locked storage locations. Evidence is checked out by investigators, and then checked back in. Information obtained from the evidence is kept in a careful manner. The crucial component of evidence use is the unbroken and complete linkage between the evidence (and the information obtained from that evidence) and the scene of the crime. If it can be argued that some other agent may have interfered with or tainted some of the evidence, this may constitute reasonable doubt, which is a key idea in the English/American process of legal argument.

While the process of scientific discovery and reasoning does not have the legal force of evidence obtained in a criminal proceeding, some of the same ideas are employed. In particular, the use of the laboratory notebook is one in which the process, methods, outcomes, and results of experiments are noted in detail (Caprette, 1995; UCSF Office of Research, 2009; UCSF Office of Technology Management, 2004). In using a lab notebook to record events as they occur in the laboratory, the scientist writes in ink for the creation of a permanent record,

dates each page, ensures that each page contains consistent information, and, in some cases, has another individual co-sign the page for additional strength of evidence. The laboratory notebook can be evidence in the determination of the granting of a patent, and thus has a level of evidentiary strength.

Tracking the process of analysis

In the era of *team science* (Bennett et al., 2010), the processes of data collection and data analysis are done by different persons. As such, there may be different approaches to the tracking of the data collection, and to the tracking of the data analysis. While statisticians and data analysis professionals have clear ideas about what constitutes a rigorous data analysis process, nonstatisticians have different ideas. In many cases, nonstatisticians perform the data analysis, and sometimes they use interactive data analysis tools. Interactive tools emphasize flexibility, as "All graphic displays benefit from being made interactive...One can control flexibly and smoothly important factors influencing the interpretation of glyph visualizations" (Gribov & Unwin, 2011). The hidden problem here lies in the postanalysis archiving of the actual analysis that "made the nice picture." The tracking process takes additional time and effort, especially if it is done after the research process has been completed. In the work of many basic scientists, the analysis of the data is not the first priority. It may be given less structure and done with less care.

If the tracking process for the analysis is initiated from the start, the entire process of analysis is performed differently. If reproducibility of the analysis process is a goal, scriptable tools rather than interactive tools will be used. Although some interactive tools have the capability of retaining the code used to perform analysis, the retention function must usually be explicitly selected during the analysis to ensure that interim work products are saved. It is often difficult to retrace steps performed using an interactive tool, unless the discipline of maintaining the scripting code is well established. In addition, if the interactive session is long and involved, the scripting code can also be long, involved, and filled with false steps and redone analyses. Thus, it is usually necessary to revisit the scripting code file to clean the code and ensure that only the successful final results are retained, along with all interim steps. It is often better to

use scripting code-based tools from the beginning to ensure that the code that is retained is the code that was desired to be retained.

Professionalism and individual integrity

Conducting and reporting research in an ethical manner demonstrates virtues of scientific professionalism including competence, openness, and accountability. It is also consistent with the notion of collaboration in the larger environment, which is an approach which produces faster progress. The careful attention required to perform research activities accurately and consistently makes scientific competence an issue not only of skill, but also of character. As observed by John Hardwig:

> Competence, conscientious work, and epistemic self-assessment are aspects of [an investigator's] "epistemic character." ... Although competence is not a character trait per se, it standardly depends on character: becoming knowledgeable and then remaining current almost always requires habits of self-discipline, focus, and persistence. (Hardwig, 1991, p. 700)

Personal habits such as self-discipline, focus, and persistence—and, we might add, honesty—operate in the background, rather than the foreground, of every research activity. They will not be mentioned in the Methods section of a scientific paper. But the strength and consistency of these habits will have a real and in some cases decisive impact on experiments, lines of research, and careers.

Reproducible Research as a Set of Ethical Scientific Techniques

The discussion of ethical behavior in science thus encompasses facets of right conduct for the individual as a collaborator, as a user of resources and subjects, and as a citizen within the larger society. Recent cases have demonstrated many ways (some discussed below) in which these ethical relationships are damaged by scientists acting both carelessly and mendaciously. Reproducible research can be seen as a response to these problems.

What is reproducible research?

Reproducible research is a way of performing scientific computations so as to ensure that the computations can be repeated by the individual later, by other individuals at any time, and by the wider scientific community involved in the publication and dissemination of scientific results. There are really three similar but disparate components to the RR movement. They are introduced here, and discussed below in greater detail in "Three Concepts of Reproducible Research."

First, reproducible research involves methods for supporting computationally intensive science. These approaches to RR involve the use of collaborative environments and computing tools which facilitate the performance, preservation, and transmission of research involving intricate computational methods.

Second, reproducible research involves methods for storing analysis of data for later reproduction. These approaches to RR involve the use of methods to preserve the computations used to support academic work products (e.g., journal articles, books, technical reports).

Third, approaches to RR involve constructing tools that will allow papers stored online, with appropriate support structures, to actually perform a task or analysis that is the subject of the paper. In standard passive documents, results are merely presented for examination. Thus, rather than simply reading about something that was done in the past, active papers allow the reader to both read about and actively perform the analysis task using the same document.

How does reproducible research address the issues of research ethics?

Reproducible research methodologies enhance the ethical behavior and activities of scientists. This relationship can be clarified in terms of the four conditions for ethical research.

Behavior of the scientist per se

In the first place, reproducible research addresses the personal and professional need to have a clear and well-defined record of the methods by which scientific results were obtained. Although in some cases investigators may be externally pressured to follow RR practices (sometimes required by sponsors, regulators, supervisors, or editorial policies), reproducibility is really a component of the personal behavior of investigators. Working to a high level of reproducibility begins with matters of technique. More importantly, it is a reflection of the virtues of honesty, self-discipline, and openness to accountability. The radical transparency of RR practices, comparable to doing science in a glass house, requires a combination of confidence in the quality of one's methods and humility in making one's work available for detailed examination and critique by others. For these reasons, RR promotes a high standard of professional conduct.

Conversely, although it may be natural enough to regard RR practices as intrusive or burdensome, especially when they are unfamiliar, investigators who persistently drag their feet on measures required for RR may actually be expressing a lack of confidence, lack of discipline, or unwillingness to facilitate review of their work by others. These attitudes give the devil a foothold where scientific integrity is concerned. Although the methods are sometimes considered to be an additional time sink with no clear benefits, disciplined use from the start of analysis considerably ameliorates the time requirements.

Working with subjects and resources

Reproducible research practices in themselves may have little or no direct impact on research subjects, who can hardly be expected to know at the time whether the research activity in which they are participating is reproducible. Nevertheless, RR shows respect for subjects and resources, as part of investigators' responsibility to conduct research in a scientifically valid manner. Research practices in which results are invalid or mischaracterized are ultimately disrespectful to subjects and sponsors.

Working with coworkers

Reproducibility is important for fulfilling the expectation within the scientific community that published scientific results should be genuinely informative and suitable for inclusion in a larger fabric of knowledge of a field of research. It may be neither possible nor desirable to actually reproduce the great majority of scientific work. But the potential ability to reproduce experiments is an important assumption behind what actually does happen on a routine basis. In science, the record of published research results provides information and inspiration for further research, adding to a web of evidence and interpretation which constitutes the evolving state of knowledge within a scientific field. Research or publication practices that misinform scientific colleagues about essential factors in the research put other scientists who work from the published results at risk of seeing their own efforts wasted or invalidated. There may also be failures of fair play within the scientific community—where competition is as much of a driving force as cooperation—when individuals or groups manipulate data analysis to give the impression of achieving a result they have in fact not achieved.

Working in society at large

The details of methodology, computation and publication that go into RR in a particular field will go far beyond the public's level of knowledge or interest. However, in the ethical sense of *the public interest* as "that which benefits the public," RR is clearly in the public interest. The connections between valid research and public good are relatively direct in fields with human applications, such as biomedicine, engineering, and environmental sciences, in which specific solutions arise as a result of research findings. But even when the connections are indirect, they are important, especially concerning the public credibility of science and technology. Public confidence in the integrity of the vast majority of individual scientists and the effectiveness of scientific peer review process is an important component in the relationship between scientists and scientific institutions and the rest of society, in particular touching on the public's willingness to embrace scientific descriptions of social problems and proposed solutions. As within the scientific community, so also in the scientific community's relationship with society as a whole, trust is paramount and may be easier to protect than to repair.

Why is reproducible research suddenly of interest?

Issues with the process of analysis and evaluation

In the current environment, and especially in light of the developments in bioinformatics and computational biology, there are a number of important and problematic trends in current research. First, there is a revolution in the amount of information. A relatively small experiment involving genetic information can easily generate 10GB of data per subject. This vast amount of information presents storage challenges as well as challenges to evaluation. It is often quite difficult to determine if the answer makes "sense" because the code is opaque, in that the relationship between data and parameter estimate is not always clear. Cluster analysis involves the analysis of (dis)similarity between entities. Once calculated, these measures are not always examined in greater detail, producing a disassociation between parameter estimate and the sets of data values. With the analysis of variance in a fixed factor situation, the variance of the dependent is fully partitioned into components, and a simple check of accuracy involves the addition of the sums of squares. With multilevel or mixed-model methods, such simple checks are not available, as the partitioning is not additive. The degrees of freedom for multilevel and mixed models are not a simple decomposition of the total N, either. Thus, contemporary computational methods often produce results which are not amenable to simple evaluations for accuracy.

Issues with the availability of tools and techniques

Modern computational technology has lowered the threshold for entry into the analysis arena. Rather than asking a statistician or data analysis professional to perform analyses, many computationally sophisticated basic scientists now perform their own analyses. This is often a good thing, but can result in problems. If a person is not a statistician or data analyst, problems are not always recognized with data analysis. These include boundary issues (problems that arise when mathematically impossible operations are performed, such as taking the square root of a negative number), issues of precision (in which values are stored in locations that are not able to store large values), and so forth. Additionally, the software which is favored by computationally less sophisticated scientists often over-promises, in that the software

advertises itself as capable of methodologies and techniques using incorrect or antiquated methods. Software is also incapable of making distinctions that statisticians make. Ordinary regression methods can be used with binary outcomes, but statisticians would use logistic regression for the analysis of this type of outcome variable.

Tracking complex operations

In the current computerized analysis era, some processes are extremely complicated by their very nature. It is a great challenge to manage, evaluate, and communicate results from these sorts of processes. For the computational disciplines (image processing, signal processing, exploratory geophysics, bioinformatics, computer science), research proceeds by the creation, modification, and application of algorithms. An algorithm is a numerical or logical technique for the manipulation of information. In some cases, an algorithm produces a value for comparison with others. In other cases, the algorithm produces a visual output (in image and signal processing in particular). Such results are often evaluated in a relatively subjective manner. Thus, when an algorithm is applied to data, the results can be difficult to evaluate. The vital problem lies in maintaining the link between results (evaluated subjectively in some cases) and input parameters.

Buckheit & Donoho (1995)

This paper discusses a number of issues and problems which led to the concepts of RR. Below are some examples (excerpt edited):

> To avoid sounding like we are pointing the finger at anyone else, we will mention a few problems we have encountered in our own research.
>
> - The Stolen Briefcase. Once, several years ago, at a conference, one of us had a briefcase stolen. ... There was no reasonable prospect of finding the time or opportunity to return ... A manuscript had already been written. The figures were so convincing ... that without them ... manuscript had to be abandoned.
> - Burning the Midnight Oil. Once, writing an article with approximately 30 figures, we had to tweak various algorithms and display options to display clearly the effects we were looking for. ... Returning to work eight hours later, we had a question: which were the "final" versions ...

- A Year is a Long Time in this Business. Once, about a year after one of us had done some work ... and ... went back to the old software library to try and (redo) it, he couldn't remember how the software worked ... he abandoned the project...
- A la Recherché des Parameters Perdues. Once, one of us read a paper ... that was very interesting ... from the paper itself he couldn't figure out what ... parameters were being used. The author ... replied, "we forgot which parameters gave the nice picture you see in the published article" ...

> For a field to qualify as a science, it is important first and foremost that published work be reproducible by others. (Buckheit & Donoho, 1995, p. 2-4)

In this case, and in many of the examples and writings by practitioners of computational sciences (signal processing, exploratory geophysics, computer sciences), the problem with conventional publication methods lies in the lack of specificity of many of the key details inherent in finished, publish work. Publication of an algorithm or computational method in a print journal usually does not include details about start values, specific parameter values for specific results, or details about the management of boundary issues (when key estimates become mathematically impossible). Sufficient detail is not found in such publications to completely, accurately, and reliably produce the same result.

In a similar manner, a discussion of situation at the Stanford Exploratory Project has these comments:

> In the mid 1980's, we noticed that a few months after completing a project, the researchers at our laboratory were usually unable to reproduce their own computational work without considerable agony. In 1991, we solved this problem by developing a concept of electronic documents that makes scientific computations reproducible. Since then, electronic reproducible documents have become our principal means of technology transfer of scientific computational research. A small set of standard commands makes a document's results and their reproduction readily accessible to any reader. To implement reproducible computational research the author must use makefiles, adhere to a community's naming conventions, and reuse (include) the community's common building and cleaning rules. Since electronic reproducible documents are reservoirs of easily maintained, reusable software, not only the reader but also the author benefits from reproducible documents. (Schwab et al, 2000)

So, in the Claerbout lab, the issue began with the idea that, within a single investigation, the results of a process should be repeatable, but that this often was truer in the omission. Schwab et al. go on to discuss passing results from one person to the next and other related topics. The key notion, again, is that the process of performing a complex analysis is difficult to communicate from one person to another. To do this, the Claerbout lab instituted the use of the makefile, which allows a series of operations to be performed in a completely consistent manner repeatedly, can easily be adapted and modified (since it is an editable file), and can be examined by a text editor to determine the key values which are used for the process.

Issues and problems in contemporary research

There have been a number of notable cases recently which illustrate some of the key notions inherent in RR.

The Potti case

In 2006, Clinical Trial TOP0602 (NCT00509366) at Duke University was launched to examine the methods of Anil Potti for the classification of cancers into groups for optimal treatment. A series of promising papers (Augustine et al., 2009; Bonnefoi et al., 2007; Hsu et al., 2007; Potti et al., 2006a; Potti et al., 2006b) had established a method of determining which drugs would most likely be effective against certain types of cancers. The screening method involved the use of microarray profiles of the NC160 (a standardized panel of cell lines derived from human tumors) in comparison with drug sensitivity data. Using this method and bioinformatic methods, the drug which was most likely to be effective against a tumor in a specific patient could be ascertained, and the patient would then be treated with this drug. The success of this method seemed quite remarkable, and oncologists at other locations became interested in duplicating and replicating the method for wider application. Two bioinformatics statisticians, K. Baggerly and K. Coombes of the MD Anderson Cancer Center, were tasked with the replication of the method at the new center. They obtained data and tools from the Potti lab at Duke and began to replicate the methods, running into considerable difficulties. Coombes and Baggerly requested additional information from Duke, and they finally concluded that trivial errors with great serious ramifications had occurred. These errors led to the termination of

three clinical trials at Duke, the termination of Anil Potti's employment, and considerable concern about the conduct of science and bioinformatic techniques.

In attempting to replicate Potti's the methods, specific steps with the data were performed that should have allowed Baggerly and Coombes to completely and fully achieve the same outcomes as did the Potti group. They were unable in many cases to do so (Baggerly & Coombes, 2009; Coombes et al., 2007; Potti & Nevins 2007). Baggerly & Coombes (2009) goes into considerable detail about the exact procedures and contains references to more voluminous and detailed records of the exact forensic bioinformatics methods that they used. The term *forensic bioinformatics* refers to the process of performing sufficient manipulation of data and technique so as to discover what actually occurred, as opposed to what was supposed to have occurred. Without going into the many details that their very interesting paper covers, a few summary observations can be made. First, the Potti group used the labels *sensitive* and *resistant* in an inconsistent manner, so that groups of patients or sets of genes were labeled as resistant at one point and sensitive at another. Second, the datasets supplied to Baggerly and Coombes had many errors, duplications of cases were found, and some of the duplicated cases were labeled inconsistently—data from one patient was labeled as both sensitive and resistant. Third, errors of labeling of genes occurred, resulting in the wrong names being placed on output, the wrong genes being identified in terms of potential members in the best.

After a lengthy and difficult discussion, the three clinical trials based on this promising methodology were cancelled, Potti was placed on leave, and Duke returned all money to the American Cancer Society that had been awarded for the clinical trials. Potti subsequently resigned his position at Duke and has recently accepted a new position at the Coastal Cancer Center, Myrtle Beach, SC, after acquiring a medical license in the state of South Carolina. For his many coworkers, it is unclear what the results of this unfortunate situation will be. At this point, the identity of the person who actually wrote the code containing the problems discussed above has not been revealed. As of October, 2011, suits have been filed naming Potti and Duke University in the deaths of some of the patients in the trial.

The research career of a promising scientist, Anil Potti, was affected and possibly ended by a combination of carelessness and bad management of data. As noted in Baggerly & Coombes (2009), several approaches can be consistently applied that would reduce or eliminate the likelihood of this sort of problem. First, a tool such as Sweave can be used to do all analyses in a reproducible manner. Second, standardized systems for study closeout, applied on a lab- or department-wide basis, can be used. Third, data coding should be done carefully; names (*Sensitive* or *Resistant*) instead of numbers (0 or 1) should be used to lend clarity to analysis. Fourth, row and column indicators should be retained with the data for all analyses, to reduce labeling errors. Fifth, results should be reviewed by colleagues. Sixth, standardized templates for statistical reports should be used for standardized situations. These approaches will not end occurrences like the Potti case, but they would structure research and analysis in a manner to make it less common.

The Baltimore case

The situation referred to by this general term offers a different set of lessons about reproducibility and scientific accountability (Baltimore, 2003; Gunsalus, 1999; Kevles, 1996; Kevles, 1998; Lang, 1994). It involved a large number of scientists, including David Baltimore, Thereza Imanishi-Kari and Margot O'Toole. David Baltimore is a noted biologist who was awarded the Nobel Prize for Physiology or Medicine in 1975 at the age of 37. In 1990, he assumed the presidency of Rockefeller University, which he resigned under pressure in 1991. Currently, he is Emeritus Professor of Biology at MIT. The events discussed here are referred to as The Baltimore Case due to the prominence and importance of David Baltimore, although he is merely one of dozens of persons involved in this case.

The reproducibility aspect of this case began with a paper (Weaver et al., 1986) of which Baltimore was a minor coauthor. After publication, the senior author, Thereza Imanishi-Kari, hired a postdoc researcher not originally involved in the paper (Margot O'Toole) to perform experiments in an attempt to extend the findings of the paper. After having difficulty repeating a key finding from the paper, O'Toole examined the original data and concluded that there were inconsistencies in the route from the data to the published paper. As noted by one critic of Baltimore, "the experimental data for that paper had been presented in a misleading

fashion" (Lang, 1994, p. 2). These doubts led her to raise concerns with others, which eventually led to involvement of the Office of Scientific Integrity (now the Office of Research Integrity) and the United States Congress House Subcommittee on Oversight and Investigations (John Dingell, D-MI 2, Chair).

The issue hinges critically upon the interpretation of data written in a lab notebook by Dr. Moema Reis, second author of the Weaver et al (1986) paper. The data in the lab notebook, in particular the "17 pages of the records" (Lang, 1994, p. 4), did not support Weaver et al. The lab notebooks were not maintained in good order, and they were sometimes not able to be found. In some cases, data were recorded out of order or on a date other than the actual date of record.

From the standpoint of reproducibility, several lessons can be drawn from this affair. First, had the original data which formed the basis for the Weaver et al. (1986) paper been archived in a reproducible manner, many of the issues of the controversy simply would not have arisen, or they would have been handled in a simple manner. A number of the contentious issues in the case involved the data. For example, were the data edited by the paper's authors? O'Toole alleged that Imanishi-Kari removed values from datasets or censored them at the time of analysis. Were values excluded to tell the "correct" story? Were certain components of key experiments even run? Second, had the data been recorded in the lab notebook in a contemporaneous, fully transparent manner, the issues involved with the forensic analysis of the lab notebooks would have been either moot or easily resolved. If the data for the paper were not present for the computations involved in several of the tables for the Weaver et al. paper, the paper could not have been written. If the data were censored or otherwise edited, as O'Toole believed had happened during an informal and private meeting between her and Imanishi-Kari, a fully specified dataset which was used in a reproducible manner to produce the paper would have revealed such editing.

Twenty-five years after the fact, it is practically impossible to ascribe complete responsibility for the issues in this complicated case, and this discussion, cursory and superficial as it is, does not attempt to ascribe blame to any party of the affair. Some claim that Baltimore was overly

zealous in his defense of Imanishi-Kari. Several of the reports about the case blame Baltimore for statements which moved the situation from the realm of science, in which open discussion, access to data, and responses to queries are the norm, to a more political area in which deference to authority is more important than response to legitimate inquiry. Some claim that Baltimore and Imanishi-Kari were unfairly treated by Dingell and the Office of Scientific Integrity. Some claim that the entire issue was politicized and became overly laden with other issues (scientific self-evaluation, the willingness of scientists to uncritically accept the judgment of highly decorated and high-status persons). The case is an important one to remember, since it illustrates many problems with the scientific establishment. It suggests again that reproducible preparation of results, and analysis can be important. With archived data, contentious issues are avoided from the start, while they can arise in the absence of well-archived data and results. While scientific fraud can be found in some cases, frequently scientific problems with important papers are the result of simple sloppiness and carelessness in handling important information. Methods that structure the presentation of results and analysis to minimize such components will lead to better outcomes.

The Chang case

The Chang case The case of Geoffrey Chang (summarized in Miller, 2006) illustrates the problems which can be caused by seemingly miniscule errors of computation. Beginning in 2001, Chang and his colleagues were able to determine the crystalline structure of several complex molecules which are important in signal transport and cell homeostasis (Chang, 2003; Chang & Roth, 2001; Ma & Chang, 2004; Pornillos et al., 2005; Reyes & Chang, 2005). Due to the size and complexity of the molecules for which the structure was able to be determined, the papers were considered highly important. A subsequent paper (Dawson & Locher, 2006) obtained different results. In reviewing the methodology of his own group, Dr. Chang determined that a "home-made data-analysis program" had "flipped two columns of data inverting the electro-density map" (Miller, 2006). This mix-up rendered all conclusions moot and forced the retraction of five papers (Chang et al., 2006).

These five initial papers had been cited a total of 676 times as of November 1, 2011, according to Web of Science Citation tracking. The retraction, published in *Science* in 2006, resulted in a

considerable decrease in citations of the papers. These papers were cited 537 times up to and including 2006, and only 129 times after. Of these 129, a number were specifically about the retraction itself, but in other cases, it is entirely unclear that the citing work is aware of the retraction. For instance, a work (Poulsen et al., 2009) that references two of the retracted works (Ma & Chang, 2004; Pornillos et al, 2005) cites these papers in a group with 5 other papers. The reference is a perfunctory one, in which the cited works merely support the notion that some components of some structures are tetrameric rather than dimeric, which would not be a conclusion compromised by the retraction. Of other papers that cited the Chang papers, it is unclear about how many of them used the Chang lab results as central, peripheral, or perfunctory components of the theoretical machinery. It is unclear how many of these papers were retracted following the Chang et al. (2006) notice of retraction. This is one of the difficult issues involved in scientific research which later is retracted. Other research is based on the retracted work(s). If the retracted work is central or key to the scientific underpinnings, this other research should be carefully evaluated.

In this case, we see again the key import of the small details involved in the computational aspects of the research. In this case, the notion of reproducibility is not entirely clear. If the program which formed the defective link did so in a consistent manner, the result would be found, regardless of the number of times that the program might be run. Thus, the reproducibility component must also include a component of second opinion or evaluative examination of the work. Check-sums, visual inspection of output, and other such mundane, seemingly irrelevant processes should always be a part of the process.

Some conclusions about fraud, malfeasance, and scientific error

These cases illustrate several important issues, which exhibit varying combinations of carelessness, recklessness, possible fraud, and scientific mismanagement. Note, however, that these are but a small sample of the many cases of problematic scientific research. The exact issue in each case is sometimes difficult to ascertain, since the key persons are often quite eager to put the difficult situations behind them and move on. Yet, in each case, issues with

data analysis, the data per se, and the exact manner in which conclusions were drawn have had wide, permanent consequences. In many cases, the exact and final conclusions about those final ramifications are unknown, since retracted articles are often cited after retraction (Budd et al., 2011), and the citations which occurred before retraction are not revisited to determine if the viability of the publication has been placed into question by the retracted publication.

Reproducible research methods, as discussed below, can be one component of the solution to these problems. By ensuring that data are available for all to examine, issues such as those in the Baltimore case might be avoided. By ensuring that computational methods are stored and examined by others, issues such as those in the Potti and Chang cases might be avoided.

Three Concepts of Reproducible Research

Reproducible research is three distinct but related streams of discussion and ideas.

Computationally intensive research support

Use of structured execution files

The term reproducible research is usually credited to Jon Claerbout, a professor of exploratory geology at Stanford University, and director of the Stanford Exploratory Project (Claerbout & Karrenbach, 1992; Schwab et al., 2000). A Claerbout quote that is often found in RR discussions is:

> An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures. (Buckheit & Donoho, 1995)

In the Stanford Exploratory Project, several issues arose over time. First, persons were unable to accurately and simply repeat their own analyses. Second, other persons (such as new graduate students) were unable to accurately and simply repeat the work of others.

For this reason, the Stanford Exploratory Project began the use of make files to encapsulate and store the processes of analysis. A *makefile* is a scripting tool for processes, common in the

Unix environment and related situations such as cygwin, the Windows system Unix emulator. This scripting tool is derived from the older .bat file system on DOS machines. In the make and .bat file approaches, commands are stored in an editable file, which is called a *structured execution file*. The commands perform operations by running programs. Looping operations, in which systematic counts are used to perform repeated operations which include a counter, can be performed. Logical operations, in which decisions can be built in to the process, can also be used, allowing the continued operation of the script to be conditional on tests of the completion of some earlier operation, when one operation depends on another. The scripted file can also be executed on a timed basis. This feature allows a dependable series of operations to be run, which produce identical outputs from identical inputs. The make system also includes standardized sets of conditions (make clean deletes interim files, make install installs software, etc.), enabling make files to perform different functions.

The make approach has limitations, however. The dependency checking capabilities of make are limited, and are based on dates, that are limited and can easily lead to difficulties. Thus, other scripting, dependency checking structured execution files have been constructed and employed for RR. One such is SCons (Fomel & Hennenfent, 2007), which is a Python-based tool for processing on data with many steps and dependent operations. It includes several important features which make computing simpler and more useful with complex data analysis situations. For instance, it includes a feature that tracks editing-produced changes and only runs code components which have been changed by editing.

Share resource environments

A shared research environment (SRE) provides research tools, shared data structures, example datasets, a systematic method for code development, and the opportunity to work in a tool environment that is also used by others in the scientific area of the user. Although the term reproducible research is used in reference to these toolsets, the emphasis is on the research component. Each component of research is reproducible when the research is embedded in an SRE. Shared research environments were developed in the computational sciences to ensure that algorithms could be developed in an environment in which comparisons between methods

could be made easily, in which continuity of technique could be assured, and in which it would relatively simple to transmit new ideas both within scientific groups and between them. This sort of methodology is widely found in computation-intensive areas (e.g., image processing, signal processing, geophysics, computer science).

Many approaches to RR were either inspired by Stanford Exploratory Project or were initiated by its graduates. These SREs include code and code development tools, datasets for testing the code, methods for communicating information about code to other users, and a group of supporters who provide help for other users. Usually these SRE efforts are open source, which means that code is furnished for minimal cost—or in most cases free—and users are encouraged to further distribute and develop the SRE. A partial list of current SRE efforts includes:

- Wavelab: used for research in the wavelet analysis area (Buckheit & Donoho, 1999). It was developed at Stanford. It is a reproducible development environment, which produces code which can be used for wavelet analysis.
- Madagascar: used for research in geophysics (Fomel, 2011; Madagascar Development Team, 2011). "The mission of the Madagascar project is to provide a shared research environment for computational data analysis in geophysics and related fields." The system provides programming tools, standard datasets, and regression tests to ensure that consistent results arise after modification.
- Clawpack: used for research in linear and nonlinear hyperbolic systems in mathematics (LeVeque, 2009). From the website: "Clawpack stands for 'Conservation Laws Package' and was initially developed for linear and nonlinear hyperbolic systems of conservation laws, with a focus on implementing high-resolution Godunov type methods using limiters in a general framework applicable to many applications."
- coNCePTuaL: a research and testing environment for network management (Pakin, 2004). This is somewhat different from some of the other efforts, in that it is a testing suite for computer networks. Yet the core notion of a library of code, which can be shared with like-minded others, is a key feature of this toolset. In that sense, it is well within the stream of the SRE.

Workflow managers

There have been a number of toolkits (Alameda et al., 2007; Bahsi et al., 2007; Castro et al., 2005; De Roure & Goble, 2009; De Roure et al., 2009; De Roure et al., 2010; Fox & Gannon, 2006; Gentleman et al., 2004; Goble et al., 2010; Hull et al., 2006; Kell & Mendes 2008; Lushbough et al., 2011; Mesirov, 2010; Reich et al., 2006; Severin et al., 2010; Tang et al., 2005) developed to manage, store, and analyze bioinformatic data. The workflow manager is a tool which allows the analyses for a process to be both managed and documented. It can run the analysis, document the analysis, and can be used as compendium to distribute the analysis. One workflow manager which is especially interesting is the Trident workflow workbench (Microsoft Research, 2008; Microsoft Research, 2009a; Microsoft Research, 2011). Trident is a workflow manager which has been used extensively in the context of weather research, oceanography, and atmospheric research (CMOP Team, 2011; LEAD Team, 2011; PS1 Science Consortium, 2011). The workflow manager is not limited to these types of applications. Rather, it is a general purpose workflow manager designed to define, administer, and run workflows. It is also designed to be used by nonprogrammers:

> Because the end users for Trident are not seasoned programmers, Trident offers a graphical interface that enables construction of workflows, launching workflows, monitoring workflows, and administration of workflows. (Microsoft Research, 2008, p. 3)

Other workflow managers are also defined to have graphical interfaces, and this does not distinguish this tool from other. Since Trident is built on a Windows platform, it may be more compatible with other Windows tools. This is a particularly interesting workflow manager for non–technically oriented scientists.

Some are beginning to address the challenge of the computer literacy barrier by enabling the work-flow manager to directly interface with MS Word (Mesirov, 2010; Microsoft Research, 2009b; Microsoft Research, 2009c), which will enable less literate users who are not willing to work with La-TeX, R, or other more advanced tools to satisfy the demands of RR. This initiative is interesting. Much of the work is being done by Microsoft Research, and thus they

have a very clear technical advantage, as they are able to get at the programmatic interface to the Word tool in a way that others are unable to match.

Biomashups

A method for workflow management which is related to but less technically sophisticated is the mashup or biomashup (Beletski, 2008; Hogan et al., 2011). Biomashups are similar to workflow management tools, but somewhat less technically challenging. They allow biologists, who are often not technically sophisticated, to define sequences of web-based services to perform tasks. The difference between a biomashup and a workflow manager is modest. Mashups are considered less sophisticated, more web-based, and are sometimes interactive in nature. Mashup environments are offered by some web developers (IBM, 2009). The IBM mashup definition environment is typical of many mashup environments, in that information for the mashup is combined through a series of wizards for data selection and templates for pages. The user can create a sophisticated and visually pleasing information representation using nothing but these tools, enabling a technically unsophisticated user to produce a final document which is complicated. The requirement for the production environment is somewhat high, since the resources must all be available to the environment, and the manners in which the information is combined must similarly be known and able to be accessed using mouse clicks and drag-and-drop methods.

Support tools for structured analysis

Many researchers do research that is not computationally intensive, but is supported by the use of statistical software. For such researchers, the analyses performed to support documents (e.g., articles, books, websites) must be preserved in a method which allows their techniques and methods to be easily and reliably reproduced. In addition, the documents themselves must be linked tightly to the analyses. In most documents, the analyses produce a set of tables and figures. The Results section contains statements such as, "As noted in Table 1, the test of Factor X is significant ($F = 10.02$, $p = .0002$)," which also need to be linked tightly to the analysis.

Statistical data analysis programs

Data analysis programs today come in a variety of types. One important distinction is that between *interactive* and *scriptable* tools. Interactive tools (e.g., jmp – SAS Institute, Inc, 2011; SPSS – IBM SPSS, Inc, 2011; GraphPad Prism – GraphPad Software, Inc, 2011) can be used with a graphical user interface (GUI) for the analysis. A GUI is a screen which has virtual buttons, drop-down menus, and other controls to allow the user to perform an analysis by a sequence of interactive choices. The results of the choices are usually displayed in windows which can then be saved into files, in the manner in which the analyst is most interested.

A scriptable tool is one in which code to perform analysis is written and executed. The code is saved in a file. The results again appear in a series of windows, and the code can either write the output directly to files, or analysis which is directed to a window can be saved. The scripted code can be edited and run. The scripts can often be generalized, so that the original data can be run, as well as other data which is similar in structure to the original data.

The two general classes of analysis are not mutually exclusive. Jmp can produce scripting files which contain jmp Scripting Language commands, a language which can then be used to perform the analysis by the submission of the scripting commands (SAS Institute, Inc., 2007). Many statistical analysis tools fall into this general class of those having an interactive capability that can also lead to scripting code which can be saved. In many cases, code-based tools can be used in a flexible manner which has the appearance of interactivity, while retaining the power and reproducibility of code-based analysis. This is done by saving work products in specific manners.

The number and variety of interactive programs seems to be growing, and this is an unfortunate consequence of the reputation of statistics as difficult. With interactive programs, tracking and documenting the actual methods of analysis are extremely difficult, and require great discipline as well as excellent documentation skills. Statisticians and professional data analysts are much less likely to use interactive tools than are nonprofessional users.

The structured analysis file

The simplest approach to reproducible research is a structured, well-commented analysis or command file. This file will contain the commands for analysis in a command-driven program (e.g., SAS, SPSS, R, Stata). This type of file, the command file, is only available when the analysis tool is a scriptable tool. The file will begin by acquiring the data from a source, proceed to prepare the data for analysis, perform the actual analysis, and conclude by preparing tables, figures, and incidental text. This is a common approach for most statisticians and data analysts, but it is not always used by basic scientists. The utility and reproducibility of this approach depends on the structure and completeness of comments and the degree to which the file has been cleaned up at the end of the analysis period. Such files can be useful, but often are not, due to the manner in which they are retained.

Literate programming tools

One key tool approach in RR is the notion of literate programming. Literate programming is a development and documentation approach to the code writing enterprise. In a literate programming effort, the code to perform operations and the documentation to describe the code are both developed together in a common system. The literate programming file is processed in one manner to produce the code for analysis and in another manner to produce the documentation. The term tangle is used to describe the process of producing the core file, and weave is used to define the process of producing documentation, code, and other products from the core. The key ideas of literate programming were developed by (Knuth, 1992) in his development of TeX, a system for mathematical and scientific typesetting. The literate programming methods have been extensively used in TeX and LaTeX. The literate programming system for LaTeX is called doc (The LaTeX3 Project, 2006). Documentation is produced by processing the .dtx file in one manner, while the working code is produced from the .dtx file by a different processing method. The first author of this paper is the author of newlfm, a LaTeX class for the production of letters, faxes, and memos, which uses the doc system (Thompson, 2008).

Literate programming concepts have been expanded and extended in the RR area. Again, the notion is to develop and manage several efforts under a single source. The several components being managed usually include the final text document, a file containing code to analyze data, a method to convert the analyses into tables and figures, and a method to combine the tables and figures with text to produce the final document. Here are a number of RR-related methods that are essentially or explicitly literate programming methods:

- The SWeave system: literate programming for the statistical system R and the LaTeX document preparation system, which links the documents produced in LaTeX with analysis results produced by R (Geyer, 2006; Leisch, 2011; Leisch, 2002a; Leisch, 2002b; Leisch, 2003; Peng & Eckel, 2009; Peng et al., 2009; Shotwell & Álvarez, 2011). From the website: "What is Sweave? Sweave is a tool that allows the user to embed the R code for complete data analyses in LaTeX documents. The purpose is to create dynamic reports, which can be updated automatically if data or analyses change. Instead of inserting a prefabricated graph or table into the report, the master document contains the R code necessary to obtain it. When run through R, all data analysis output (tables, graphs, etc.) is created on the fly and inserted into a final LaTeX document. The report can be automatically updated if data or analysis change, which allows for truly reproducible research."
- The STATWEAVE system: literate programming for statistically oriented document preparation (Lenth, 2008). From the website: "STATWEAVE is software whereby you can embed statistical code (e.g., SAS, R, Stata, etc.) into a LaTeX or OpenOffice document. After running STATWEAVE, the code listing, output, and graphs are added to the document." STATWEAVE grew out of SASWeave (Lenth & Højsgaard, 2007), and has essentially replaced it.
- Attempts are beginning to be made address the problem of WYSIWYG text editing tools such as MS Windows. Workflow managers which directly interface with MS Word (Mesirov 2010) enable less literate users (who are not willing to work with LaTeX, R, or other advanced tools) to satisfy the demands of RR. This approach is limited since it is derived and based on a specific workflow manager (GenePattern), but is important in

that this inclusive approach attempts to bring less sophisticated users to the RR concept.

Research compendia

The research compendium (Gentleman, 2005; Gentleman & Temple Lang, 2003; Gentleman & Temple Lang, 2007) is another reproducibility tool:

> In its most simplistic form a compendium is a collection of software, data and one or more navigable documents. A navigable document is a document that contains markup for both text chunks and code chunks. It is called a navigable document because a reader, equipped with the appropriate software tools, can navigate its contents and explore and reproduce the computationally derived content. (Gentleman, 2005, p. 2).

This idea combines the extended notion of literate programming in the Buckheit & Donoho sense with the notion of having all information in a container that can then be easily and simply given to others.

Inherent in the notion of the compendium is the notion of a shared library of computational tools. If there is a common substrate of computational tools, it is necessary to include only the specific code to use that tool. If the tool is not commonly found in the community of users, the tool itself would be included as part of the compendium. Naturally, this would be somewhat more difficult because tools run in specific environments; Windows-based tools are different from Linux-based tools. Until 1990, statistical analysis users could reliably expect that other statistical analysis users would have either SAS or SPSS as a tool. Since 1990, R/S+ has come to be the tool of choice for statisticians, bioinformaticians, and epidemiologists, although SAS retains a strong position. Stata has also come to be a tool of choice for many applied statisticians. In this sense, the shared research environment is the method of distribution and computational support when a common tool is not available or when the tool is itself the interest of the research. The compendium is more useful when it can be assumed that a common tool is available and will be available for those who are attempting to replicate and understand the research available in the compendium.

Active document support

In the last several years, publishers are becoming more aware of RR issues. In 2010, Elsevier Publishing issued the "Executable Paper Grand Challenge" (Elsevier Ltd., 2010), which sought solutions to the problems faced by publishers in the area of executable papers. From the initial press release: "Elsevier, a leading global publisher of scientific, technical and medical information, announced today that it has launched a new competition that invites scientists to improve how data intensive research is represented in a scholarly journal article. ... Winning ideas should have innovative answers to the question, how can data intensive research be made repeatable and workable within the context of a scholarly journal article?" (Elsevier Ltd., 2010). The results were presented at the International Conference on Computational Science 2011 (ICCS 2011, June 1-3, 2011, Singapore). The first place winner was "The COLLAGE authoring environment" (Nowakowski et al., 2011), which represents the document as a combination of static text and active code. Data is also embedded in the document. The second place paper was "SHARE: a web portal for creating and sharing executable research papers"(Gorpa & Mazanek, 2011). This paper presents the idea of a shared virtual machine, which would support processing required by a paper. Authors would provide the mechanism, and readers could then use the mechanism, adapting it to their own data if appropriate. The third place paper was "A Universal Identifier for Computational Results" (Gavish & Donoho, 2011). The remaining papers from the competition (Brammer et al., 2011; Gomes et al., 2011; Hinson, 2011; Jourjon et al., 2011; Kauppinen & de Espindola, 2011; Kohlhase et al., 2011; Leisch, 2011; Limare & Morel, 2011; Müller et al., 2011; Siciarek & Wiszniewski, 2011; Strijkers et al., 2011; Veres & Adolfsson, 2011) presented a variety of approaches, many of which proposed a web-accessible database for paper support, in which authors would have editing access (enabling them to change things permanently), while readers would have reading and manipulation access (enabling them to manipulate the results).

Reproducible Research Initiatives in Contemporary Science

Reproducible research has made inroads into specific scientific areas at a much greater degree than in other areas. Scientific areas which have made great contributions and show great involvement with the notions of RR are listed here.

Biostatics/Statistics/Epidemiology/Bioinformatics (listed together due to overlap in interests, personnel, and programs):

- Reproducibility focus: reproducible computation, active documents
- Research groups: Biostatistics, Johns Hopkins University; Statistics, Stanford University; Bioinformatics and Computational Biology, University of Texas M.D. Anderson Cancer Center; Biostatistics, Vanderbilt University; MADA Center, Sanford Research/USD; Statistics, University of Iowa; Statistics, University of California–Los Angeles
- Discussions in the literature:Potti situation (Baggerly & Coombes; 2009, Baggerly & Coombes, 2011; Coombes et al., 2007); compendia and other packaging tools (Gentleman, 2005; Gentleman & Temple Lang, 2003, Gentleman & Temple Lang, 2007); documentation requirements for clinical research (Baggerly et al., 2004a; Baggerly et al., 2004b; Liedtke et al., 2010; Ochs & Casagrande, 2008); documentation and presentation requirements (Hothorn et al., 2009; Peng & Eckel, 2009; Peng et al., 2006; Peng et al., 2009; Peng, 2011a; Peng, 2011b)
- Shared Research Environment tools: Wavelab (wavelet analysis), workflow managers, compendia

Computational Biology:

- Reproducibility focus: computationally intensive research, reproducible computation, active documents

Econometrics:

- Reproducibility focus: computationally intensive research, reproducible computation, active documents
- Research groups: Economics, University of Illinois; Economics, Fordham University
- Discussions in the literature: computationally intensive research (Kroenker & Zeileis, 2009), output support methods (McCullough & Vinod, 2003; Vinod, 2000; Vinod, 2001)

Geophysics:

- Reproducibility focus: computationally intensive algorithm development. Research groups: Stanford Exploratory Project, Stanford University (Palo Alto); Department of Geophysics, University of Texas (Houston)
- Discussions in the literature: reproducibility (Fomel & Hennenfent, 2007); tool use (Fomel, 2011; Madagascar Development Team, 2011)
- Shared Research Environment tools: Madagascar

Image Processing/Computer Vision (overlap in personnel and foci):

- Research focus: computationally intensive algorithm development
- Research groups: Computer Science and Electrical Engineering, West Virginia University; Electrical and Computer Engineering & Biomedical Engineering, Carnegie Mellon University

Psychology:

- Discussions in the literature: data monitoring and archiving (Stevens, 2010)

Signal Processing

- Research focus: computationally intensive algorithm development
- Research groups: Electrical and Computer Engineering & Biomedical Engineering Carnegie Mellon University; Electrical Engineering and Computer Science, University of California–Berkeley
- Discussions in the literature: reproducibility and science (Barni & Perez-Gonzalez, 2005; Barni et al., 2007; Cook & Vandewalle, 2011); promoting reproducibility (Kóvacévic, 2007); experience with reproducibility (Vandewalle et al., 2007a; Vandewalle, 2011a; Vandewalle et al., 2007b; Vandewalle et al., 2009; Vandewalle 2011b)

The Future of Reproducibility

The reproducible research proposals and active methods have already had an impact on research and publication methods in specific areas (discussed above in "Reproducible research initiatives in contemporary science").

New subject areas

Many scientific areas have yet to embark upon the discussion of reproducibility. These include many of the social sciences, many engineering disciplines, areas of mathematics, and quantitative history. The reproducibility challenges involved in mathematics—in which automated derivation and proof tools would be needed to verify proofs—would be considerably different from the requirements of sociology or psychology. However, these issues will be coming to all sciences.

The quantity of research today

Science in the whole is being inundated by research results. Every day, it seems like a new journal is founded. Libraries are being required to subscribe to hundreds of journals. In addition to the many journals published by traditional publishers, a vast number of open-access journals are being founded. For example, six new journals, published and supported by the Public Library of Science (PLoS), have been founded in the last eight years. These are PLoS: Biology, Genetics, Neglected Tropical Diseases, Medicine, Pathogens, Programming Languages and Operating Systems, and Computational Biology. These are widely respected and influential journals, and they are not available in print format.

The sheer amount of information being published simply emphasizes that the entire issue of ensuring that the quality of science is a very difficult one. If it took a single hour to perform a reproducibility analysis on every article being published in the Journal of the American Medical Association in one year, it would take 400 hours, or the ten weeks' work for a single, full-time statistician (the figure of 1500 hours was noted for the forensic bioinformatics work documented in Baggerly Coombes, 2009). Since there are more than 2000 medical journals, the amount of time it would take to perform full reproducibility analyses on all such journals is large.

Pros and cons of technological solutions

The reproducible research notion is based on highly integrated and strongly linked relationships between data, programs, and documents. As time goes by, this level of

integration can cause serious problems. Computer programs are similar to living things, as they are being changed all the time. In the best of possible worlds, they maintain backward compatibility, but this is not always possible. Tools change, and sometimes the changes are fundamental. In addition, the interface to the tool may be consistent, but the code being used to perform some operation may change. Thus, although the user may be running the "same" operation or analysis, the analysis may in fact be different in several ways. The start values may be determined in a different manner. The stop values may also be different. If the procedure is iterative and asymptotically convergent upon an answer, the criteria for final convergence are quite central to the solution of the analysis. If these are changed, the solution will often be quite different.

In addition, the location of information on the web can frequently change. Recent studies suggest that large proportions of web addresses become inoperable, unstable, or unavailable over time (Dellavalle et al., 2003; Koehler 2004; Markwell & Brooks, 2003; Markwell & Brooks, 2008; Rhodes, 2010; Thorp & Brown, 2007; Wagner, 2009), a phenomena that is sometimes called link rot (Markwell & Brooks, 2003; Rhodes, 2010) or URL decay (Wren, 2008). The entire address of an institution changes at times, and sometimes the entire system of addresses will need to change. The current system of address storage is an agreement, not a guarantee, and it may change over time. The URL system will need to be expanded in the near future, based on the rate at which the addressing capacity is being depleted. Some approaches, such as the DOI system, attempt to provide permanent addresses, but these are not universally used the system itself is new, and thus may change over time.

Many operative concepts of reproducibility are dependent on capacity and resources to maintain of the current structure of resources. Since many of the programs being used for RR are open-source, and thus based on the willingness of volunteers to maintain the code, this is a shaky edifice at the very best. As the amount of code, data, and documents expand at the rate that they are currently expanding, the "dead weight" of the past information will become large. Will it be sustainable? It is not uncommon to find inactive links in the pages of currently active scientists. What will happen when these scientists are no longer active?

The use of catalogues of services is one approach which might be of value (Bhagat et al., 2010). A catalogue, such as the BioCatalogue, is a web-based list of services available for biological evaluation and analysis. The BioCatalogue contains information about services, annotations about the services (e.g., descriptions of tools and information about tool use), and information about the information required to run a service as well as the output from the service. This is a promising development, and might be long-term a solution to a number of problems.

Reproducible Research in the Methodology and Data Analysis Center

In the Methodology and Data Analysis Center, all documents are expected to be at RR Level 2 (see **Table 1**). As documents are completed, the full script is archived in a consistent structure. All datasets are retained in final analysis form, all script files are saved, and all tables and figures are retained as well. The Sanford Research document tool is Microsoft Word, and thus moving to RR Level 3 will take some additional tools. Active methods are under construction to add this capability.

All archives have a consistent structure:

Name-subject-year

Data
- Excel
- Raw
- SAS

Documents
- Figures
- References
- Statistical results

Programs

Within the Programs subdirectory, the main analysis program is called *master.sas*. This file contains a shell analysis macro (%macro _overall), preset inclusion statements for files containing other macros (macros.sas), formats (formats.sas), and preset macros to output results to correct places in the structure. In this way, a new project begins by copying the master pattern structure, adding a single reference to the subdirectory root Name-subject-year, and correctly defined coding can then begin. In this approach, RR is also research which can be easily and successfully inherited from analysts who leave, and which can be worked on collaboratively by several analysts within the Methodology and Data Analysis Center.

Table 1: Levels of Reproducible Research

Level 1	Scripts are used to run analysis
Level 2	Tables and figures are produced entirely by script Incidentals added to paper by hand
Level 3	Tables and figures inserted into paper by script Incidental values inserted into paper by script
Level 4	Paper is built by same script as does analysis Entire document is produced by same script
Level 5	Data and program are included with paper in publication

Description: Documents can be defined at several levels of reproducibility. Each level builds upon the previous levels. As increasing RR levels are attained, the process becomes more consistent, and the work product can be more strongly relied upon.

Conclusion

Reproducible research is a somewhat new idea, in that the term first came into the common parlance of science about 25 years ago. Yet, it is not new either. It is really the oldest idea in science that scientists are working to attempt to determine facts about reality which in turn can be relied upon as truth. In this way, RR is not simply an interesting idea, but is an essential manner of doing science in a modern world of increasing complexity.

References

Alameda, J., Christie, M., Fox, G., Futrelle, J. Gannon, D., Hategan, M. ...Thomas, M. (2007). The open grid computing environments collaboration: portlets and services for science gateways. *Concurrency and Computation: Practice & Experience, 19*(6), 921–942. doi: 10.1002/cpe.1078

Augustine, C.K., Yoo, J.S., Yoshimoto, Y., Zipfel, P.A., Friedman, H.S., Nevins, J.R....Tyler, D.S. (2009). Genomic and molecular profiling predicts response to temozolomide in melanoma. *Clinical Cancer Research, 15*, 502–510.

Baggerly, K., Morris, J., & Coombes, K. (2004). High-resolution serum proteomic patterns for ovarian cancer detection. *Endocrine-Related Cancer, 11*, 583–584.

Baggerly, K., Morris, J., & Coombes, K. (2004). Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics and Statistical Applications of Genetics and Molecular Biology, 20*, 777–785.

Baggerly, K.A. & Coombes, K.R. (2009). Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Annals of Applied Statistics, 3*(4), 1309–1334. doi: 10.1214/09-AOAS291

Baggerly, K.A. & Coombes, K.R. (2011). What information should be required to support clinical "omics" publications? *Clinical Chemistry, 57*(5), 688–690. doi: 10.1373/clinchem.2010.158618

Bahsi, E.M., Ceyhan, E., & Kosar, T. Conditional workflow management: A survey and analysis. *Scientific Programming, 15*(4), 283–297.

Baltimore, D. (2003). Baltimore's travels. *Issues in Science and Technology, 19*, 1.

Barni, M. & Perez-Gonzalez, F. (2005). Pushing science into signal processing. *IEEE Signal Processing Magazine, 22*(4), 119–120.

Barni, M., Perez-Gonzalez, F., Comesaña, P. & Bartoli, G. (2007) Putting reproducible signal processing into practice: A case study in watermarking. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 4*, 1261–1264.

Beletski, O. (2008). *End user mashup programming environments* (Technical Report T-111). Helskinki: Helsinki University of Technology Telecommunications Software and Multimedia Laboratory.

Bennett, L.M., Gadlin, H, & Levine-Finley, S. (2010). *Collaboration and team science: A field guide.* Washington, DC: National Institutes of Health.

Bezabeh, T., Evelhoch, J.L., Sloop, D.J., Thompson, P.A., & Ackerman, J.J.H. (2004). Therapeutic efficacy as predicted by quantitative assessment of murine RIF1 tumor pH and phosphorus metabolite response during hyperthermia: An in-vivo 31 PNMR study. *International Journal of Hyperthermia, 20,* 335–357.

Bhagat, J., Tanoh, F., Nzuobontane, E., Laurent, T., Orlowski, J. , Roos, M., ...Goble, C.A.. (2010). Biocatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acids Research, 38*, W689–W694. doi: 10.1093/nar/gkq394

Bonnefoi, H., Potti, A., Delorenzi, M., Mauriac, L., Campone, M., Tubiana-Hulin, M....Iggo, R.D. (2007). Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the eortc 10994/big 00-01 clinical trial. *Lancet: Oncology, 8*(12), 1071–1078. doi: 10.1016/S1470-2045(07)70345-5

Brammer, G.R., Crosby, R.W., Matthews, S.J., & Williams, T.L. (2011). Paper mache: Creating dynamic reproducible science. *Procedia Computer Science: International Conference on Computational Science, ICCS 2011, 4*, 658–667.

Buckheit, J.B., & Donoho, D.L. (1995). WaveLab and reproducible research. In Antoniadis, A. & Oppenheim, G. (Eds.), *Lecture Notes in Statistics Series: Wavelets in Statistics*, (pp. 55–82). New York: Springer-Verlag.

Buckheit, J.B. & Donoho, K. (1999). *Wavelab and reproducible research* (Technical Report 474). Stanford: Stanford University Department of Statistics. Retrieved from http://statistics.stanford.edu/~ckirby/techreports/NSF/EFS%20NSF%20474.pdf

Budd, J.M., Coble, Z.C., & Anderson, K.M. (2011). *Retracted publications in biomedicine: Cause for concern.* Paper presented at Association of College and Research Libraries Annual Conference, Philadelphia, Pennsylvania, March 30-April 2, 2011.

Caprette, D.R. (1995). Guidelines for keeping a laboratory record. Retrieved from http://www.ruf.rice.edu/~bioslabs/tools/notebook/notebook.html

Castro, A.G., Thoraval, S., Garcia, L.J., & Ragan, M.A. (2005). Workflows in bioinformatics: Meta-analysis and prototype implementation of a workflow generator. *BMC Bioinformatics, 6*, 87. doi: 10.1186/1471-2105-6-87

Chang, G. (2003). Structure of msba from vibrio cholera: A multidrug resistance abc transporter homolog in a closed conformation. *Journal of Molecular Biology, 330*(2), 419–430. doi: 10.1016/S0022-2836(03)00587-4

Chang, G. & Roth, C.B. (2001). Structure of msba from e-coli: A homolog of the multidrug resistance atp binding cassette (abc) transporters [retracted article. see v. 314, p. 1875, 2006]. *Science, 293*(5536), 1793–1800. doi: 10.1126/science.293.5536.1793

Chang, G., Roth, C.B., Reyes, C.L., Pornillos, O., Chen, Y-J., & Chen, A.P. (2006). Retraction. *Science, 314*, 1875.

Claerbout, J.F. & Karrenbach, M. (1992). Electronic documents give reproducible research a new meaning. *Proceedings of the 62nd Annual International Meeting of the Society of Exploration Geophysics*, 601–604.

CMOP Team. (2011). Center for coastal margin observation & prediction (cmop). Retrieved from http://www.stccmop.org/

Cook, J. & Vandewalle, P. (2011). Reproducible research ideas. Retrieved from goo.gl/fjeme

Coombes, K.R., Wang J., & Baggerly, K.A. (2007). Microarrays: retracing steps [Letter to the editor]. *Nature Medicine, 13*, 1276–1277. doi: 10.1038/nm1107-1276b

Dawson, R.J.P., & Locher, K.P. (2006). Structure of a bacterial multidrug ABC transporter. Nature, 443 (7108), 180–185. doi: 10.1038/nature05155

De Roure, D. & Goble, C. (2009). Software design for empowering scientists. IEEE Software, 26 (1), 88–95.

De Roure, D., Goble, C., & Stevens, R. (2009). The design and realisation of the (my)Experiment Virtual Research Environment for social sharing of workflows. *Future Generation Computer Systems, 25*(5), 561–567. doi: 10.1016/j.future.2008.06.010

De Roure, D., Goble,C., Aleksejevs, A., Bechhofer, S., Bhagat, J., Cruickshank, D., ... Poschen, P. (2010). Towards open science: the myExperiment approach. *Concurrency and Computation: Practice & Experience, 22*(17, SI), 2335–2353. doi: 10.1002/cpe.1601

Dellavalle, R.P., Hester, E.J., Heilig, L.F., Drake, A.L., Kuntzman, J.W., Graber, M., & Schilling, L.M. (2003). Going, going, gone: Lost internet references. *Science, 302,* 787–788.

Elsevier, Ltd. Executable paper grand challenge. (2010). Retrieved from http://www.executablepapers.com/

Fomel, S. (2011). Reproducible research: Lessons from the MADAGASCAR project [presentation]. *SIAM Conference on Data Mining (SDM11)*. Retrieved from http://jarrodmillman.com/talks/siam2011/ms148/fomel.pdf

Fomel, S. & Hennenfent, G. (2007). Reproducible computational experiments using SCons.Volume 4, 1257–1260.

Fox, G.C. & Gannon, D. (2006). Special issue: Workflow in grid systems. *Concurrency and Computation: Practice & Experience, 18*(10), 1009–1019. doi: 10.1002/cpe.1019

Gavish, M., & Donoho, D. (2011). A universal identifier for computational results. *Procedia Computer Science: International Conference on Computational Science, ICCS 2011, 4*, 637–647.

Gentleman, R. (2005). Reproducible research: A bioinformatics case study. *Statistical Applications in Genetics and Molecular Biology, 4*(1). doi: 10.2202/1544-6115.1034

Gentleman, R. & Temple Lang,D. (2003). *Statistical analyses and reproducible research* (Technical Report 2).

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S. ... Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology, 5*(10), R80. Retrieved from http://genomebiology.com/2004/5/10/R80

Gentleman, R. & Temple Lang, D. (2007). Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics, 16*(1), 1–23.

Geyer, C. An Sweave demo, 2006. Retrieved from http://users.stat.umn.edu/~geyer/Sweave/

Goble, C.A., Bhagat, J., Aleksejevs, S., Cruickshank, D., Michaelides, D., Newman, D., ... David De Roure. (2010). myExperiment: a repository and social network for the sharing of bioinformatics workflows. *Nucleic Acids Research, 38*, W677–W682.

Gomes, A.T.A., Paredes, D., & Valentin, F. (2011). Supporting the perpetuation and reproducibility of numerical method publications. *Procedia Computer Science: International Conference on Computational Science, ICCS 2011, 4*, 688–696.

Gorpa, P. Van, & Mazanek, S. (2011). SHARE: a web portal for creating and sharing executable research papers. *Procedia Computer Science: International Conference on Computational Science, ICCS 2011, 4*, 589–597.

GraphPad Software, Inc. (2011). GraphPad Prism. Retrieved from http://graphpad.com

Gribov, A. & Unwin, A. (2011). What is Gauguin? Retrieved from http://rosuda.org/software/Gauguin/gauguin.html

Gunsalus, C.K. (1999). Book reviews: "The Baltimore case: A Trial of Politics, Science, and Character" by D. Kevles. *The New England Journal of Medicine, 340,* 242.

Hardwig, J. (1991). The role of trust in knowledge. *Journal of Philosophy, 88*, 693–708.

Hinson, K. (2011). A data and code model for reproducible research and executable papers. *Procedia Computer Science: International Conference on Computational Science, ICCS 2011, 4*, 579–588.

Hogan, J.M., Sumitomo, J., Roe, P., & Newell, F. (2011). Biomashups: the new world of exploratory bioinformatics? *Concurrency and Computation: Practice and Experience, 23*(11), 1169–1178. doi: 10.1002/cpe.1598

Hothorn, T., Held, L., & Friede, T. (2009). Biometrical journal and reproducible research. *Biometrical Journal, 51*, 553–555.

Hsu, D.S., Balakumaran, B.S., Acharya, C.R., Vlahovic, V., Walters, K.S., Garman, K. ... Potti, A. (2007). Pharmacogenomic strategies provide a rational approach to the treatment of cisplatin-resistant patients with advanced cancer. *Journal of Clinical Oncology, 25*, 4350–4357.

Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M.R., Li, P., & Oinn, T. (2006). Taverna: a tool for building and running workflows of services. *Nucleic Acids Research, 34*, W729–W732.

IBM. (2009). Lotus mashups. Retrieved from goo.gl/nN5ue

IBM SPSS, Inc. (2011). IBM SPSS Statistics. Retrieved from http://www-01.ibm.com/software/analytics/spss/products/statistics/

Jourjon, G., Rakotoarivelo, T., Dwertmann, C., & Ott, M. (2011). LabWiki : An executable paper platform for experiment-based research. *Procedia Computer Science: International Conference on Computational Science, ICCS 2011, 4*, 697–706.

Kauppinen, T., & Mira de Espindola, G. (2011). Linked open science? Communicating, sharing and evaluating data, methods and results for executable papers. *Procedia Computer Science: International Conference on Computational Science, ICCS 2011, 4*, 726–731.

Kell, D.B., & Mendes, P. (2008). The markup is the model: Reasoning about systems biology models in the semantic web era. *Journal of Theoretical Biology, 252*(3), 538–543. doi: 10.1016/j.jtbi.2007.10.023

Kevles, D. (1996, May 27). Annals of science: The assault on David Baltimore. *The New Yorker*, p. 80.

Kevles, D. (1998). *The Baltimore Case: A trial of politics, science, and character*. New York: W.W. Norton & Co.

Knuth, D.E. (1992). Literate Programming. *CSLI Lecture Notes* (Vol. 27). Stanford, CA: Center for the Study of Language and Information.

Koehler, W. (2004). A longitudinal study of web pages continued: a consideration of document persistence. *Information Research: An International Electronic Journal, 9*(2).

Kohlhase, M., Corneli, J., David, C., Ginev, D., Jucovschi, C., Kohlhase, A.,..... Zholudev, V. (2011). The planetary system: Web 3.0 & active documents for STEM. *Procedia Computer Science: International Conference on Computational Science, ICCS 2011, 4*, 598–607.

Kóvacévic, J. (2007) How to encourage and publish reproducible research. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 4*, 1273–1276.

Kroenker, R., & Zeileis, A. (2009). On reproducible econometric research. *Journal of Applied Econometrics, 24*(5), 836–847. doi: 10.1002/jae.1083

Laine, C., Goodman, S.N., Griswold, M.E., & Sox, H.C. (2007). Reproducible research: Moving toward research the public can really trust. *Annals of Internal Medicine, 146*, 450–453.

Lang, S. (1994). Questions of scientific responsibility: The Baltimore case. *Ethics and Behavior, 3*, 3–72.

LEAD Team. (2011). Linked environments for atmospheric discovery ii (lead ii). Retrieved from http://d2i.indiana.edu/leadII-home

Leisch, F. (2011). Executable papers for the r community: The r 2 platform for reproducible research. *Procedia Computer Science: International Conference on Computational Science, ICCS 2011, 4*, 618–626.

Leisch, F. (2002). Sweave, part I: Mixing R and latex. *R News, 2*(3): 28–31. Retrieved from http://cran.r-project.org/doc/Rnews/

Leisch, F. (2002). Dynamic generation of statistical reports using literate data analysis. In Härdle, W., & B Rönz, B. (Eds). *COMPSTAT 2002: Proceedings in Computational Statistics* (pp. 575-580). Heidelberg: Physica Verlag.

Leisch, F. (2003). Sweave, part II: Package vignettes. *R News, 3*(2), 21–24. Retrieved from http://cran.r-project.org/doc/Rnews/

Lenth, R.W. (2008). StatWeave – Beta version. Retrieved from http://homepage.stat.uiowa.edu/~rlenth/StatWeave/

Lenth, R.V., & Højsgaard, S. (2007). SASweave: Literate programming using SAS. *Journal of Statistical Software, 19* (8), 1–20. Retrieved from http://www.jstatsoft.org/v19/i08/

Lenze, E.J., Goate, A.M., Nowotny, P., Dixon, D., Shi, P., Bies, R.B.... Bruce G. Pollock.(2010). Relation of serotonin transporter genetic variation to efficacy of escitalopram for generalized anxiety disorder in older adults. *Journal of Clinical Psychopharmacology, 30*(6), 672–677. doi: 10.1097/JCP.0b013e3181fc2bef

LeVeque, R.J. (2009). Python tools for reproducible research on hyperbolic problems. *Computing in Science and Engineering, 11*, 28–34. doi: 10.1109/MCSE.2009.13

Liedtke, C., Wang, J., Tordai, A., Symmans, W.F., Hortobagyi, G.N., Kiesel, L. ... Pusztai, L. (2010). Clinical evaluation of chemotherapy response predictors developed from breast cancer cell lines. *Breast Cancer Research And Treatment, 121*(2), 301–309. doi: 10.1007/s10549-009-0445-7

Limare, N. & Morel, J-M. (2011).The IPOL initiative: Publishing and testing algorithms on line for reproducible research in image processing. *Procedia Computer Science: International Conference on Computational Science, ICCS 2011, 4*, 716–725.

Lushbough, C.M., Jennewein, D.M., & Brendel, V.P. (2011). The BioExtract Server: a web-based bioinformatic workflow platform. *Nucleic Acids Research, 39*(Supp. 2), W528–W532. doi: 10.1093/nar/gkr286

Ma, C., & Chang, G. (2004). Structure of the multidrug resistance efflux transporter EmrE from Escherichia coli. *Proceedings of The National Academy Of Sciences Of The United States Of America, 101*(9), 2852–2857. doi: 10.1073/pnas.0400137101

Madagascar Development Team. (2011). Madagascar. Retrieved from http://reproducibility.org/wiki/Package_overview

Markwell, J. & Brooks, D.W. (2003). "Link rot" limits the usefulness of web-based educational materials in biochemistry and molecular biology. *Biochemistry and Molecular Biology Education, 31*(1), 69–72.

Markwell, J., &. Brooks, D.W. (2008). Evaluating web-based information: Access and accuracy. *Journal of Chemical Education, 85*(3), 458–459.

McCullough, B.D., & Vinod, H.D. (2003). Verifying the solution from a nonlinear solver: A case study. *American Economic Review, 93*, 873–892.

Mesirov, J.P. (2010). Accessible reproducible research. *Science, 327*, 415–416.

Microsoft Research. (2008). *Trident workbench: A scientific workflow system* (technical report). Microsoft Corporation. Retrieved from http://research.microsoft.com/en-us/collaboration/tools/trident.aspx

Microsoft Research. (2009).Trident composer user's guide [white paper]. Microsoft Corporation.

Microsoft Research. (2009). *Trident document add-in for microsoft word: User guide* (technical report). Microsoft Corporation. Retrieved from http://research.microsoft.com/en-us/collaboration/tools/trident.aspx

Microsoft Research. (2009). *Trident workflow word add-in for Microsoft Word: Onboarding guide* (technical report). Microsoft Corporation. Retrieved from http://research.microsoft.com/en-us/collaboration/tools/trident.aspx

Microsoft Research. (2011). Project trident: A scientific workflow work. Retrieved from http://tridentworkflow.codeplex.com/documentation

Miller, G. (2006). A scientist's nightmare: Software problem leads to five retractions. *Science, 314*, 1856–1857.

Müller, M., Rojas, I., Eberhart, A., Haase, P., & Schmidt, M. (2011). A-R-E: The author-review-execute environment. *Procedia Computer Science: International Conference on Computational Science, ICCS 2011, 4*, 627–636.

Nowakowski,P., Ciepiela, E., Harzlak, D., Kocot, J., Kasztelnik, M., Bartynski,T.,... Malawski, M. (2011). The Collage Authoring Environment. *Procedia Computer Science: International Conference on Computational Science, ICCS 2011, 4*, 608–617.

Ochs, M.F., & Casagrande, J.T. (2008). Information systems for cancer research. *Cancer Investigation, 26*(10), 1060–1067. doi: 10.1080/07357900802272729

Pakin, S. (2004). Reproducible network benchmarks with coNCePTuaL. In Danelutto, M., Vanneschi, M. & Laforenza, D. (Eds.) *Proceedings of Euro-Par 2004 Parllel Processing, 10th International Euro-Par Conference, Pisa, Italy, August 31-September 3, 2004* (pp. 64–71).

Peng, R.D., & Eckel, S.P. (2009). Distributed reproducible research using cached computations. *Computing in Science and Engineering, 11*(1), 28–34. doi: 10.1109/MCSE.2009.6

Peng, R.D., Dominici, F., Zeger, S.L. (2006). Reproducible epidemiologic research [commentary]. *American Journal of Epidemiology, 163*(9), 783–789. doi: 10.1093/aje/kwj093

Peng, R.D., Diggle, P.J., Zeger, S.L. (2009). Reproducible research and Biostatistics. *Biostatistics, 10*(3), 405–408.

Peng, R.D. (2011). Computational and policy tools for reproducible research [presentation]. *Reproducible Research: Tools and Strategies for Scientific Computing, a workshop at Applied Mathematics Perspectives 2011, an ICIAM Satellite meeting held at the University of British Columbia in Vancouver, BC, Canada, July 13-16, 2011.*

Peng, R.D. (2011). Reproducible research in computational science. *Science, 334*, 1226–1227. doi: 10.1126/science.1213847

Pornillos, O., Chen, Y.J., Chen, A.P. (2005). X-ray structure of the emre multidrug transporter in complex with a substratei. *Science, 310*(5756), 1950–1953. doi: 10.1126/science.1119776

Potti, A., Dressman, H.K., Bild, A., Riedel, R.F., Chan, G., Sayer, R. ... Nevins, J.R. (2006). Genomic signatures to guide the use of chemotherapeutics. *Nature Medicine, 12*(11), 1294–1300. doi: 10.1038/nm1491

Potti, A., Mukherjee, S., Petersen, R., Dressman, H.K., Bild, A., Koontz, J. ... Nevins, J.R. (2006). A genomic strategy to refine prognosis in early-stage non–small-cell lung cancer. *New England Journal of Medicine, 355*(6), 570–580.

Potti, A. & Nevins, J.R. (2007). Reply to "Microarrays: Tracing steps." *Nature Medicine, 13*, 1277–1278.

Poulsen, B.E., Rath, A., & Deber, C.M. (2009).The assembly motif of a bacterial small multidrug resistance protein. *Journal of Biological Chemistry, 284*, 9870–9875.

PS1 Science Consortium. (2011). Pan-starrs (panoramic survey telescope & rapid response system). Retrieved from http://pan-starrs.ifa.hawaii.edu/public/home.html

Reich, M., Liefeld, T., Gould, J., Lerner,J., Tamayo, P, & Mesirov, J.P. (2006). Genepattern 2. *Nature Genetics, 38*, 500–501.

Reyes, C.L & Chang, G. (2005). Structure of the ABC transporter MsbA in complex with ADP-vanadate and lipopolysaccharide. *Science, 308*(5724), 1028–1031. doi: 10.1126/science.1107733

Rhodes, S. (2010). Breaking down link rot: The Chesapeake Project Legal Information Archive's examination of URL stability. *Law Library Journal, 102*(4), 581–597.

SAS Institute, Inc. (2007). *JMP Scripting Guide, Release 7* (technical report). Retrieved from http://www.jmp.com/support/downloads/pdf/jmp_scripting_guide.pdf

SAS Institute, Inc. (2011). JMP. Retrieved from http://www.jmp.com/software/

SAS Institute, Inc. (2011). SAS Version 9.3. Retrieved from http://www.sas.com/software/sas9/

Schwab, M., Karrenbach, M., & Claerbout, J.F. (2000). Making scientific computations reproducible. *Computing in Science and Engineering, 2*, 61–67.

Severin, J., Beal, K., Vilella, A.J., Fitzgerald, S., Schuster, M., Gordon, L. ... Herrero, J. (2010). ehive: An artificial intelligence workflow system for genomic analysis. *BMC Bioinformatics, 11*. doi: 10.1186/1471-2105-11-240

Shotwell, M.S. & Álvarez, J.M. (2011, July 1). Approaches and barriers to reproducible practices in biostatistics [presentation]. Vanderbilt University. Retrieved from http://biostatmatt.com/uploads/shotwell-interface-2011.pdf

Siciarek, J. & Wiszniewski, B. (2011). IODA - an interactive open document architecture. *Procedia Computer Science: International Conference on Computational Science, ICCS 2011, 4*, 668–677.

Stevens, J.R. (2010). The challenges of understanding animal minds [Specialty Grand Challenge Article]. *Frontiers in Psychology*. doi: 10.3389/fpsyg.2010.00203

Strijkers, R., Cushing, R., Vasyunin, D., de Laat, C., Belloum, A.S.Z., & Meijer, R. (2011). Toward executable scientific publications. *Procedia Computer Science: International Conference on Computational Science, ICCS 2011, 4*, 707–715.

Tang, F., Chua, C.L., Ho, L-Y., Lim, Y.P., Issac, P., & Krishnan, A. (2005). Wildfire: distributed, grid-enabled workflow construction and execution. BMC Bioinformatics, 6, 69. doi: 10.1186/1471-2105-6-69

The LaTeX3 Project. (2006). Latex2e for class and package writers.

Thompson, P.A. (2009, April 11) newlfm: A class for letters, faxes, and memos for Latex2e. Retrieved from http://ctan.unixbrain.com/macros/latex/contrib/ne wlfm/manual.pdf

Thorp, A.W., & Brown, L. (2007). Accessibility of internet references in annals of emergency medicine: Is it time to require archiving. *Annals of Emergency Medicine, 50*(2), 188–192.

UCSF Office of Research. (2009). Guidelines for lab notebooks. Retrieved from http://www.research.ucsf.edu/qg/orQgNb.asp

UCSF Office of Technology Management. (2004). OTM recommended laboratory procedures. Retrieved from http://www.otm.ucsf.edu/docs/otmLabProc.asp

Vandewalle, J., Suykens, J., de Moor, B., & Lendasse, A. (2007). State of the art and evolutions in public data sets and competitions for system identification, time series prediction and pattern recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 4*, 1269–1272.

Vandewalle, P. (2011). Experiences with reproducible research in various facets of signal processing research.

Vandewalle, P., Barrenetxea, G., Jovanovic,I., Ridolfi, A., & Vetterli, M. (2007). Experiences with reproducible research in various facets of signal processing research. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 4*, 1253–1256.

Vandewalle, P., Kovacevic, J., & Vetterli, M. (2009). Reproducible research in signal processing — what, why, and how. *IEEE Signal Processing Magazine, 26*(3), 37–47.

Vandewalle, P. (2011). Reproducible research in signal processing: How to increase impact [presentation]. *Applied Mathematics Perspectives 2011, an ICIAM Satellite meeting held at the University of British Columbia in Vancouver, BC, Canada, July 13-16, 2011.* Retrieved from http://www.stodden.net/AMP2011/slides/reproducible_research_Vancouver_2011_07_14-1.pdf

Veres, S.M., & Adolfsson, J.P. (2011). A natural language programming solution for executable papers. *Procedia Computer Science: International Conference on Computational Science, ICCS 2011, 4*, 678–687.

Vinod, H.D. (2000). Review of GAUSS for Windows, including its numerical accuracy. *Journal of Applied Econometrics, 15*, 211–220.

Vinod, H.D. (2001). Care and feeding of reproducible econometrics. *Journal of Econometrics, 100*, 87–88.

Wagner, C., Gebremichael, M.D., Taylor, M.K., & Soltys, M.J. (2009). Disappearing act: decay of uniform resource locators in health care management journals. *Journal of the Medical Library Association, 97*(2), 122–130. doi: 10.3163/1536-5050.97.2.009

Weaver, D., Reis, M.H., Albanese, C., Costantini, F., Baltimore, D., & Imanishi-Kari, T. (1986). Altered repertoire of endogenous immunoglobulin gene expression in transgenic mice containing a rearranged Mu heavy chain gene. *Cell, 45*(2), 247–259.

Wren, J.D. (2008). URL decay in MEDLINE — a 4-year follow-up study. *Bioinformatics, 24*(11), 1381–1385.

ncpre

National Center for Professional and Research Ethics