

Best Practices for Scientific Computing

Greg Wilson ^{*}, D.A. Aruliah [†], C. Titus Brown [‡], Neil P. Chue Hong [§], Matt Davis [¶], Richard T. Guy ^{||}, Steven H.D. Haddock ^{**}, Katy Huff ^{††}, Ian M. Mitchell ^{‡‡}, Mark D. Plumbley ^{§§}, Ben Waugh ^{¶¶}, Ethan P. White ^{***}, Paul Wilson ^{†††}

^{*}Software Carpentry (gwwilson@software-carpentry.org), [†]University of Ontario Institute of Technology (Dhavid.Aruliah@uoit.ca), [‡]Michigan State University (ctb@msu.edu), [§]Software Sustainability Institute (N.ChueHong@epcc.ed.ac.uk), [¶]Space Telescope Science Institute (mrdavis@stsci.edu), ^{||}University of Toronto (guy@cs.utoronto.ca), ^{**}Monterey Bay Aquarium Research Institute (steve@practicalcomputing.org), ^{††}University of Wisconsin (khuff@cae.wisc.edu), ^{‡‡}University of British Columbia (mitchell@cs.ubc.ca), ^{§§}Queen Mary University of London (mark.plumbley@eecs.qmul.ac.uk), ^{¶¶}University College London (b.waugh@ucl.ac.uk), ^{***}Utah State University (ethan@weecology.org), and ^{†††}University of Wisconsin (wilsonp@engr.wisc.edu)

Scientists spend an increasing amount of time building and using software. However, most scientists are never taught how to do this efficiently. As a result, many are unaware of tools and practices that would allow them to write more reliable and maintainable code with less effort. We describe a set of best practices for scientific software development that have solid foundations in research and experience, and that improve scientists' productivity and the reliability of their software.

Software is as important to modern scientific research as telescopes and test tubes. From groups that work exclusively on computational problems, to traditional laboratory and field scientists, more and more of the daily operation of science revolves around computers. This includes the development of new algorithms, managing and analyzing the large amounts of data that are generated in single research projects, and combining disparate datasets to assess synthetic problems.

Scientists typically develop their own software for these purposes because doing so requires substantial domain-specific knowledge. As a result, recent studies have found that scientists typically spend 30% or more of their time developing software [19, 52]. However, 90% or more of them are primarily self-taught [19, 52], and therefore lack exposure to basic software development practices such as writing maintainable code, using version control and issue trackers, code reviews, unit testing, and task automation.

We believe that software is just another kind of experimental apparatus [63] and should be built, checked, and used as carefully as any physical apparatus. However, while most scientists are careful to validate their laboratory and field equipment, most do not know how reliable their software is [21, 20]. This can lead to serious errors impacting the central conclusions of published research [43]: recent high-profile retractions, technical comments, and corrections because of errors in computational methods include papers in *Science* [6], *PNAS* [39], the *Journal of Molecular Biology* [5], *Ecology Letters* [37, 8], the *Journal of Mammalogy* [33], and *Hypertension* [26].

In addition, because software is often used for more than a single project, and is often reused by other scientists, computing errors can have disproportional impacts on the scientific process. This type of cascading impact caused several prominent retractions when an error from another group's code was not discovered until after publication [43]. As with bench experiments, not everything must be done to the most exacting standards; however, scientists need to be aware of best practices both to improve their own approaches and for reviewing computational work by others.

This paper describes a set of practices that are easy to adopt and have proven effective in many research settings. Our recommendations are based on several decades of collective experience both building scientific software and teaching computing to scientists [1, 65], reports from many other groups [22, 29, 30, 35, 41, 50, 51], guidelines for commercial

and open source software development [61, 14], and on empirical studies of scientific computing [4, 31, 59, 57] and software development in general (summarized in [48]). None of these practices will guarantee efficient, error-free software development, but used in concert they will reduce the number of errors in scientific software, make it easier to reuse, and save the authors of the software time and effort that can be used for focusing on the underlying scientific questions.

1. Write programs for people, not computers.

Scientists writing software need to write code that both executes correctly and can be easily read and understood by other programmers (especially the author's future self). If software cannot be easily read and understood it is much more difficult to know that it is actually doing what it is intended to do. To be productive, software developers must therefore take several aspects of human cognition into account: in particular, that human working memory is limited, human pattern matching abilities are finely tuned, and human attention span is short [2, 23, 38, 3, 55].

First, a program should not require its readers to hold more than a handful of facts in memory at once (1.1). Human working memory can hold only a handful of items at a time, where each item is either a single fact or a "chunk" aggregating several facts [2, 23], so programs should limit the total number of items to be remembered to accomplish a task. The primary way to accomplish this is to break programs up into easily understood functions, each of which conducts a single, easily understood, task. This serves to make each piece of the program easier to understand in the same way that breaking up a scientific paper using sections and paragraphs makes it easier to read. For example, a function to calculate the area of a rectangle can be written to take four separate coordinates:

```
def rect_area(x1, y1, x2, y2):
    ...calculation...
```

or to take two points:

```
def rect_area(point1, point2):
    ...calculation...
```

The latter function is significantly easier for people to read and remember, while the former is likely to lead to errors, not

Reserved for Publication Footnotes

least because it is possible to call it with values in the wrong order:

```
surface = rect_area(x1, x2, y1, y2)
```

Second, *names should be consistent, distinctive, and meaningful (1.2)*. For example, using non-descriptive names, like `a` and `foo`, or names that are very similar, like `results` and `results2`, is likely to cause confusion.

Third, *code style and formatting should be consistent (1.3)*. If different parts of a scientific paper used different formatting and capitalization, it would make that paper more difficult to read. Likewise, if different parts of a program are indented differently, or if programmers mix `CamelCaseNaming` and `pothole_case_naming`, code takes longer to read and readers make more mistakes [38, 3].

Finally, where possible, *all aspects of software development should be broken down into tasks roughly an hour long (1.4)*, and programmers should pace themselves to maximize long-term productivity. Our brains get tired: on short time scales, focus starts to fade after an hour to ninety minutes of intense concentration, and there is a sharp increase in error rates. In practice, this means working in chunks of 50–200 lines of code at a time [53, 7]. On longer time scales, total productivity is maximized when people work roughly 40 hours a week because error rates increase markedly past that point [55].

2. Automate repetitive tasks.

Computers were invented to do these kinds of repetitive tasks but, even today, many scientists type the same commands in over and over again or click the same buttons repeatedly [1]. In addition to wasting time, sooner or later even the most careful researcher will lose focus while doing this and make mistakes.

In practice, scientists should *rely on the computer to repeat tasks (2.1)* and *save recent commands in a file for re-use (2.2)*. For example, most command-line tools have a “history” option that lets users display and re-execute recent commands, with minor edits to filenames or parameters. This is often cited as one reason command-line interfaces remain popular [54, 18]: “do this again” saves time and reduces errors.

A file containing commands for an interactive system is often called a *script*, though in practice there is no difference between this and a program. The Unix shell, and the Python, R, and MATLAB interpreters all make it easy for users to experiment with commands, then create a record after the fact of exactly what they did to produce a particular result. As we will discuss in Section 10, this also aids reproducibility.

When these scripts are repeatedly used in the same way, or in combination, a workflow management tool can be used. The paradigmatic example is compiling and linking programs in languages such as Fortran, C++, Java, and C# [11]. The most widely used tool for this task is probably `Make*`, although many alternatives are now available [60]. All of these allow people to express dependencies between files, i.e., to say that if A or B has changed, then C needs to be updated using a specific set of commands. These tools have been successfully adopted for scientific workflows as well [15].

To avoid errors and inefficiencies from repeating commands manually, scientists should *use a build tool to automate their scientific workflows (2.3)*, e.g., specify the ways in which intermediate data files and final results depend on each other, and on the programs that create them, so that a single command will regenerate anything that needs to be regenerated.

3. Use the computer to record history.

Careful record keeping is fundamental to science of all kinds. Just as lab notebooks are considered crucial to document work at the bench, it is important to have a detailed record of the data manipulation and calculations that have been performed using computers. Therefore, *software tools should be used to track computational work automatically (3.1)*, allowing each step to be captured accurately.

In order to maximize reproducibility, everything needed to re-create the output should be recorded automatically in a format that other programs can read. (Borrowing a term from archaeology and forensics, this is often called the *provenance* of data.) There have been some initiatives to automate the collection of this information, and standardize its format [47], but it is already possible to record the following without additional tools:

- unique identifiers and version numbers for raw data records (which scientists may need to create themselves);
- unique identifiers and version numbers for programs and libraries;
- the values of parameters used to generate any given output; and
- the names and version numbers of programs (however small) used to generate those outputs.

In practice, many results are produced by interactive exploration. In these cases, the interpreter’s “history” command can be used to save recent commands to a file as a record of how particular results were produced. Such files are often used as starting points for writing scripts to automate future work (Section 2).

4. Make incremental changes.

Unlike traditional commercial software developers, but very much like developers in open source projects or startups, scientific programmers usually don’t get their requirements from customers, and their requirements are rarely frozen [57, 58]. In fact, scientists often *can’t* know what their programs should do next until the current version has produced some results. This challenges design approaches that rely on specifying requirements in advance.

Many software development teams now believe that programmers should *work in small steps with frequent feedback and course correction (4.1)* rather than trying to plan months or years of work in advance. While the details vary from team to team, these developers typically work in steps that are sized to be about an hour long, and these steps are often grouped in iterations that last roughly one week. This accommodates the cognitive constraints discussed in Section 1, and acknowledges the reality that real-world requirements are constantly changing. The goal is to produce working (but incomplete) code after each iteration. While these practices have been around for decades, they gained prominence starting in the late 1990s under the banner of *agile development* [40, 36].

5. Use version control.

Two of the biggest challenges scientists and other programmers face when working with code and data are keeping track of changes (and being able to revert them if things go wrong), and collaborating on a program or dataset [41]. Typical “solutions” are to email software to colleagues or to copy succes-

*<http://www.gnu.org/software/make>

†<http://www.dropbox.com>

sive versions of it to a shared folder, e.g., Dropbox[‡]. However, both approaches are fragile and can lead to confusion and lost work when important changes are overwritten or out-of-date files are used. It's also difficult to find out which changes are in which versions or to say exactly how particular results were computed at a later date.

The standard solution in both industry and open source is to use a *version control system* (5.1) (VCS) [42, 14]. A VCS stores snapshots of a project's files in a *repository* (or a set of repositories). Programmers can modify their working copy of the project at will, then *commit* changes to the repository when they are satisfied with the results to share them with colleagues.

Crucially, if several people have edited files simultaneously, the VCS highlights the differences and requires them to resolve any conflicts before accepting the changes. The VCS also stores the entire history of those files, allowing arbitrary versions to be retrieved and compared, together with meta-data such as comments on what was changed and the author of the changes. All of this information can be extracted to provide provenance for both code and data.

Many good VCSes are open source and freely available, including Subversion[‡], Git[§], and Mercurial[¶]. Many free hosting services are available as well (SourceForge^{||}, Google Code^{**}, GitHub^{††}, and BitBucket^{‡‡} being the most popular). As with coding style, the best one to use is almost always whatever your colleagues are already using [14].

In practice, *everything that has been created manually should be put in version control* (5.2), including programs, original field observations, and the source files for papers. Automated output and intermediate files can be regenerated at need. Binary files (e.g., images and audio clips) may be stored in version control, but it is often more sensible to use an archiving system for them, and store the metadata describing their contents in version control instead [45].

6. Don't repeat yourself (or others).

Anything that is repeated in two or more places is more difficult to maintain. Every time a change or correction is made, multiple locations must be updated, which increases the chance of errors and inconsistencies. To avoid this programmers follow the DRY Principle [25], for “don't repeat yourself”, which applies to both data and code.

For data, this maxim holds that *every piece of data must have a single authoritative representation in the system* (6.1). For example, physical constants should be defined exactly once to ensure that the entire program is using the same value, and raw data files should have a single canonical version. Similarly, every location where data was collected should be recorded once and given an ID. Every observation from that site should then include that ID instead of duplicating the site's latitude and longitude.

The DRY Principle applies to code at two scales. At small scales, *code should be modularized rather than copied and pasted* (6.2). Avoiding “code clones” has been shown to reduce error rates [28]: when a change is made or a bug is fixed, that change or fix takes effect everywhere, and people's mental model of the program (i.e., their belief that “this one's been fixed”) remains accurate. As a side effect, modularizing code allows people to remember its functionality as a single mental chunk, which in turn makes code easier to understand. Modularized code can also be more easily repurposed for other projects.

At large scales, scientific programmers should *re-use code instead of rewriting it* (6.3). Tens of millions of lines of high-quality open source software are freely available on the web,

and at least as much is available commercially. It is typically better to find an established library or package that solves a problem than to attempt to write one's own routines for well established problems (e.g., numerical integration, matrix inversions, etc.).

7. Plan for mistakes.

Mistakes are inevitable, so verifying and maintaining the validity of code over time is immensely challenging [17]. While no single practice has been shown to catch or prevent all mistakes, several are very effective when used in combination [42, 10, 56].

Defensive programming. The first line of defense is *defensive programming*: programmers should *add assertions to programs to check their operation* (7.1). An *assertion* is simply a statement that something holds true at a particular point in a program; as the example below shows, assertions can be used to ensure that inputs are valid, outputs are consistent, and so on^{§§}.

```
def bradford_transfer(grid, point, smoothing):
    assert grid.contains(point),
        'Point is not located in grid'
    assert grid.is_local_maximum(point),
        'Point is not a local maximum in grid'
    assert len(smoothing) > FILTER_LENGTH,
        'Not enough smoothing parameters'
    ...do calculations...
    assert 0.0 < result <= 1.0,
        'Bradford transfer value out of legal range'
    return result
```

Assertions can make up a sizeable fraction of the code in well-written applications, just as tools for calibrating scientific instruments can make up a sizeable fraction of the equipment in a lab. These assertions serve two purposes. First, they ensure that if something does go wrong, the program will halt immediately, which simplifies debugging. (Few things are as frustrating as slowly working backward from a crash or a wrong answer to try to find its root cause.)

Second, assertions are *executable documentation*, i.e., they explain the program as well as checking its behavior. This makes them more useful in many cases than comments since the reader can be sure that they are accurate and up to date.

Write and run tests. The second layer of defense is *automated testing*. Automated tests can check to make sure that a single unit of code is returning correct results, or check to make sure that the behavior of a program doesn't change when the details are modified. These tests are conducted by the computer, so they can be rerun every time the program is modified, to make sure that the changes have not accidentally introduced bugs.

The core of testing is the *unit test*, which checks the correctness of a single unit of software, which typically means a single function or method. Unit tests are the building blocks of any quality assurance effort: after all, if the components in

[‡]<http://subversion.apache.org>

[§]<http://git-scm.com>

[¶]<http://mercurial.selenic.com>

^{||}<http://sourceforge.net>

^{**}<http://code.google.com>

^{††}<https://github.com>

^{‡‡}<https://bitbucket.org>

^{§§}Assertions do not require language support: it is common in languages such as Fortran for programmers to create their own test-and-fail functions for this purpose.

a program are unreliable, the program isn't likely to be reliable either. Larger scale *integration testing* check that pieces of code work correctly when combined; in scientific computing, this is often done by comparing output to experimental data or the results of earlier programs that are trusted.

At either scale, *regression testing* is the practice of running pre-existing tests after changes to the code in order to make sure that it hasn't regressed, i.e., that things which were working haven't been broken. By providing this feedback, regression testing gives programmers confidence that the changes they're making are actually progress. Every project should therefore strive to make regression testing easy, so that programmers will actually do it.

In order to manage their tests, programmers should *use an off-the-shelf unit testing library (7.2)* to initialize inputs, run tests, and report their results in a uniform way. These libraries are available for all major programming languages, including Fortran, C/C++, IDL, MATLAB, Python, R, and others commonly used in scientific computing [66, 44, 49]. Exactly *how* they check correctness depends on the researcher's understanding of the problem at hand [24, 32, 46]. What the tests accomplish is automatically checking to see whether the code matches the researcher's expectations of its behavior. As such, good automated testing improves our confidence that the code is operating properly, and that the results it produces are valid.

One significant benefit of adopting a testing library is that it encourages programmers to design and build code that is testable. In practice, this means creating self-contained functions and classes with well-defined interfaces that can run more or less independently of one another. Code that is designed this way is also easier to understand (Section 1) and more reusable (Section 6).

If a group wants to start testing software that *hasn't* been built this way, the first step is to refactor legacy code to make it testable; i.e., reorganize or rewrite that code in ways that do not change its behavior [16, 34] but which may result in less tightly coupled chunks. Since smaller units of code usually have simpler behavior, refactoring legacy code is often a matter of breaking functions, classes, or modules into smaller, more testable pieces. This can be done incrementally and systematically [13], e.g., by introducing testing in "high uncertainty" areas or as old algorithms are replaced with new ones.

Use a variety of oracles. An *oracle* is something which tells a developer how a program should behave or what its output should be. In commercial software development, the oracle is often a contract or specification written by a business specialist. In scientific research, oracles include analytic results (e.g., closed-form solutions to special cases or simplified versions of the problem), experimental results, and results produced by earlier, trusted, programs. These can all provide useful checks, so programmers should *use all available oracles when testing programs (7.3)*.

Turn bugs into test cases. No matter how carefully software is tested, some bugs will inevitably sneak by and need to be fixed. In order to prevent those bugs from reappearing, programmers should *turn bugs into test cases (7.4)*, by writing tests that trigger the bug and (once fixed) will prevent the bug from reappearing unnoticed. Doing this is one way to build up a suite of regression tests, particularly for legacy programs.

Use a symbolic debugger. Having admitted that a few bugs will always make it past our defenses, our next recommendation is that programmers should *use a symbolic debugger (7.5)*

to track them down. A better name for this kind of tool would be "interactive program inspector" since a debugger allows users to pause a program at any line (or when some condition is true), inspect the values of variables, and walk up and down active function calls to figure out why things are behaving the way they are.

Debuggers are usually more productive than adding and removing print statements or scrolling through hundreds of lines of log output [67], because they allow the user to see exactly how the code is executing rather than just snapshots of state of the program at a few moments in time. In other words, the debugger allows the scientist to witness what is going wrong directly, rather than having to anticipate the error or infer the problem using indirect evidence.

One practice we *don't* advocate, even though many of us rely on it, is *test-driven development (TDD)*. When using TDD, the programmer writes the test cases for a new piece of code before writing the code itself. This may seem backward, but writing the tests helps the programmer clarify the purpose of the code in her own mind (i.e., it serves as a design aid), and also helps ensure that tests actually get written.

The reason we don't advocate it is that a meta-study of its effectiveness done in 2010 did not find any significant impact on programmer productivity [62]. Some of us interpret this to mean that we don't really know how to measure the productivity of programmers, but unless and until other results emerge, we are obliged as scientists to label TDD as "interesting if true".

8. Optimize software only after it works correctly.

Today's computers and software are so complex that even experts find it hard to predict which parts of any particular program will be performance bottlenecks [27]. The most productive way to make code fast is therefore to make it work correctly, determine whether it's actually worth speeding it up, and—in those cases where it is—to *use a profiler to identify bottlenecks (8.1)*.

This strategy also has interesting implications for choice of programming language. Research has confirmed that most programmers write roughly the same number of lines of code per unit time regardless of the language they use [53]. Since faster, lower level, languages require more lines of code to accomplish the same task, scientists should *write code in the highest-level language possible (8.2)*, and shift to low-level languages like C and Fortran only when they are sure the performance boost is needed^{¶¶}. Taking this approach allows more code to be written (and tested) in the same amount of time. Even when it is known before coding begins that a low-level language will ultimately be necessary, rapid prototyping in a high-level language helps programmers make and evaluate design decisions quickly. Programmers can also use a high-level prototype as a test oracle for a high-performance low-level reimplementation, i.e., compare the output of the optimized (and usually more complex) program against the output from its unoptimized (but usually simpler) predecessor in order to check its correctness.

9. Document design and purpose, not mechanics.

In the same way that a well documented experimental protocol makes research methods easier to reproduce, good documentation helps people understand code. This makes the

^{¶¶}Using higher-level languages also helps program comprehensibility, since such languages have, in a sense, "pre-chunked" the facts that programmers need to have in short-term memory

code more reusable and lowers maintenance costs [42]. As a result, code that is well documented makes it easier to transition when the graduate students and postdocs who have been writing code in a lab transition to the next career phase. Reference documentation and descriptions of design decisions are key for improving the understandability of code. However, inline documentation that recapitulates code is *not* useful. Therefore we recommend that scientific programmers *document interfaces and reasons, not implementations (9.1)*. For example, a clear description at the beginning of a function that describes what it does and its inputs and outputs is useful, whereas the comment in the code fragment below does nothing to aid comprehension:

```
i = i + 1      # Increment the variable 'i' by one.
```

If a piece of code requires substantial description of the implementation to be understandable, it is generally recommended that one *refactor code instead of explaining how it works (9.2)*, i.e., rather than write a paragraph to explain a complex piece of code, reorganize the code itself so that it doesn't need such an explanation. This may not always be possible—some pieces of code simply are intrinsically difficult—but the onus should always be on the author to convince his or her peers of that.

The best way to create and maintain reference documentation is to *embed the documentation for a piece of software in that software (9.3)*. Doing this increases the probability that when programmers change the code, they will update the documentation at the same time.

Embedded documentation usually takes the form of specially-formatted and placed comments. Typically, a *documentation generator* such as Javadoc, Doxygen^{***} extracts these comments and generates well-formatted web pages and other human-friendly documents.

10. Collaborate.

In the same way that having manuscripts reviewed by other scientists can reduce errors and make research easier to understand, reviews of source code can eliminate bugs and improve readability. A large body of research has shown that *code reviews* are the most cost-effective way of finding bugs in code [12, 7]. They are also a good way to spread knowledge and good practices around a team. In projects with shifting membership, such as most academic labs, code reviews help ensure that critical knowledge isn't lost when a student or postdoc leaves the lab.

Code can be reviewed either before or after it has been committed to a shared version control repository. Experience shows that if reviews don't have to be done in order to get code into the repository, they will soon not be done at all [14]. We therefore recommend that projects *use pre-merge code reviews (10.1)*.

An extreme form of code review is *pair programming*, in which two developers sit together while writing code. One (the driver) actually writes the code; the other (the navigator)

provides real-time feedback and is free to track larger issues of design and consistency. Several studies have found that pair programming improves productivity [64], but many programmers find it intrusive. We therefore recommend that teams should *use pair programming when bringing someone new up to speed and when tackling particularly tricky problems (10.2)*.

Once a team grows beyond a certain size, it becomes difficult to keep track of what needs to be reviewed, or of who's doing what. Teams should therefore *use an issue tracking tool (10.3)* to maintain a list of tasks to be performed and bugs to be fixed [9]. This helps avoid duplicated work and makes it easier for tasks to be transferred to different people. Free repository hosting services like GitHub include issue tracking tools, and many good standalone tools exist as well, such as Trac^{†††}.

Conclusion

We have outlined a series of recommended best practices for scientific computing based on extensive research, as well as our collective experience. These practices can be applied to individual work as readily as group work; separately and together, they improve the productivity of scientific programming and the reliability of the resulting code, and therefore the speed with which we produce results and our confidence in them. They are also, we believe, prerequisites for reproducible computational research: if software is not version controlled, readable, and tested, the chances of its authors being able to re-create results (much less anyone else) are remote.

Research suggests that the time cost of implementing these kinds of tools and approaches in scientific computing is almost immediately offset by the gains in productivity of the programmers involved [1]. Even so, the recommendations described above may seem intimidating to implement. Fortunately, the different practices reinforce and support one another, so the effort required is less than the sum of adding each component separately. Nevertheless, we do not recommend that research groups attempt to implement all of these recommendations at once, but instead suggest that these tools be introduced incrementally over a period of time.

How to implement the recommended practices can be learned from many excellent tutorials available online or through workshops and classes organized by groups like Software Carpentry^{†††}. This type of training has proven effective at driving adoption of these tools in scientific settings [1].

Computing is now central to the practice of science. For this aspect of scientific research to obtain the level of rigor that is applied throughout the scientific process, it is necessary for scientists to begin to adopt the tools and approaches that are known to improve both the quality of software and the efficiency with which it is produced. Doing so will improve our confidence in the results of computational science and will allow us to make rapid progress on important scientific questions that would otherwise not be possible.

1. Jorge Aranda. Software Carpentry Assessment Report, 2012.
2. Alan Baddeley, Michael W. Eysenck, and Michael C. Anderson. Memory. Psychology Press, 2009.
3. D. Binkley, M. Davis, D. Lawrie, and C. Morrell. To CamelCase or Under_score. In 2009 IEEE International Conference on Program Comprehension, 2009.
4. Jeffrey C. Carver, Richard P. Kendall, Susan E. Squires, and Douglass E. Post. Software Development Environments for Scientific and Engineering Software: A Series of Case Studies. In 29th International Conference on Software Engineering, 2007.
5. Geoffrey Chang. Retraction of 'Structure of MsbA from Vibrio cholera: A Multidrug Resistance ABC Transporter Homolog in a Closed Conformation' [J. Mol. Biol. (2003) 330 419430]. Journal of Molecular Biology, 369(2), 2007.

6. Geoffrey Chang, Christopher B. Roth, Christopher L. Reyes, Owen Pornillos, Yen-Ju Chen, and Andy P. Chen. Retraction. Science, 314(5807):1875, 2006.

7. Jason Cohen. Modern Code Review. In Andy Oram and Greg Wilson, editors, Making Software: What Really Works, and Why We Believe It, pages 329–336. O'Reilly, 2010.

^{***}http://en.wikipedia.org/wiki/Comparison_of_documentation_generators

^{†††}<http://trac.edgewall.org>

^{†††}<http://software-carpentry.org>

8. David Currie and Jeremy Kerr. Testing, as opposed to supporting, the Mid-domain Hypothesis: a response to Lees and Colwell (2007). *Ecology Letters*, 10(9):E9–E10, 2007.
9. P. Dubois and J. Johnson. Issue Tracking. *Computing in Science & Engineering*, 5(6), November-December 2003.
10. P. F. Dubois. Maintaining Correctness in Scientific Programs. *Computing in Science & Engineering*, 7(3):80–85, May-June 2005.
11. P. F. Dubois, T. Epperly, and G. Kumfert. Why Johnny Can't Build (Portable Scientific Software). *Computing in Science & Engineering*, 5(5):83–88, 2003.
12. Michael E. Fagan. Design and Code Inspections to Reduce Errors in Program Development. *IBM Systems Journal*, 15(3), 1976.
13. Michael Feathers. *Working Effectively with Legacy Code*. Prentice Hall, 2004.
14. Karl Fogel. *Producing Open Source Software: How to Run a Successful Free Software Project*. O'Reilly, 2005.
15. S. Fomel and G. Hennenfent. Reproducible computational experiments using SCons. In *32nd International Conference on Acoustics, Speech, and Signal Processing*, 2007.
16. Martin J. Fowler. *Refactoring: Improving the Design of Existing Code*. Addison-Wesley, 1999.
17. Penny Grubb and Armstrong A. Takang. *Software Maintenance: Concepts and Practice*. World Scientific, 2 edition, 2003.
18. Steven Haddock and Casey Dunn. *Practical Computing for Biologists*. Sinauer Associates, 2010.
19. Jo Erskine Hannay, Hans Petter Langtangen, Carolyn MacLeod, Dietmar Pfahl, Janice Singer, and Greg Wilson. How Do Scientists Develop and Use Scientific Software? In *Second International Workshop on Software Engineering for Computational Science and Engineering*, 2009.
20. L. Hatton. The T Experiments: Errors in Scientific Software. *Computational Science & Engineering*, 4(2):27–38, 1997.
21. L. Hatton and A. Roberts. How Accurate is Scientific Software? *IEEE Transactions on Software Engineering*, 20(10):785–797, 1994.
22. Michael A. Heroux and James M. Willenbring. Barely-Sufficient Software Engineering: 10 Practices to Improve Your CSE Software. In *Second International Workshop on Software Engineering for Computational Science and Engineering*, 2009.
23. Roger R. Hock. *Forty Studies That Changed Psychology: Explorations into the History of Psychological Research*. Prentice Hall, 6th edition, 2008.
24. Daniel Hook and Diane Kelly. Testing for Trustworthiness in Scientific Software. In *Second International Workshop on Software Engineering for Computational Science and Engineering*, May 2009.
25. Andrew Hunt and David Thomas. *The Pragmatic Programmer: From Journeyman to Master*. Addison-Wesley, 1999.
26. Hypertension. Notice of Retraction. *Hypertension*, 2012.
27. Michael B. Jones and John Regehr. The Problems You're Having May Not Be the Problems You Think You're Having: Results from a Latency Study of Windows NT. In *7th Workshop on Hot Topics in Operating Systems*, 1999.
28. Elmar Juergens, Florian Deissenboeck, Benjamin Hummel, and Stefan Wagner. Do Code Clones Matter? In *31st International Conference on Software Engineering*, 2009.
29. David Kane. Introducing Agile Development into Bioinformatics: An Experience Report. In *Agile Development Conference 2005*, 2005.
30. David Kane, Moses Hohman, Ethan Cerami, Michael McCormick, Karl Kuhlman, and Jeff Byrd. *Agile Methods in Biomedical Software Development: a Multi-Site Experience Report*. *BMC Bioinformatics*, 7(1):273, 2006.
31. Diane Kelly, Daniel Hook, and Rebecca Sanders. Five Recommended Practices for Computational Scientists Who Write Software. *Computing in Science & Engineering*, 11(5):48–53, 2009.
32. Diane Kelly and Rebecca Sanders. Assessing the Quality of Scientific Software. In *First International Workshop on Software Engineering for Computational Science and Engineering*, May 2008.
33. Douglas A. Kelt, James A. Wilson, Eddy S. Konno, Jessica D. Braswell, and Douglas Deutschman. Differential Responses of Two Species of Kangaroo Rat (*Dipodomys*) to Heavy Rains: A Humbling Reappraisal. *Journal of Mammalogy*, 89(1):252–254, 2008.
34. Joshua Kerievsky. *Refactoring to Patterns*. Addison-Wesley, 2004.
35. Sarah Killcoyne and John Boyle. Managing Chaos: Lessons Learned Developing Software in the Life Sciences. *Computing in Science & Engineering*, 11(6):20–29, 2009.
36. Henrik Kniberg. *Scrum and XP from the Trenches*. Lulu.com, 2007.
37. David C. Lees and Robert K. Colwell. A strong Madagascan rainforest MDE and no equatorward increase in species richness: re-analysis of 'The missing Madagascan mid-domain effect', by Kerr J.T., Perring M. & Currie D.J. (*Ecology Letters* 9:149159, 2006). *Ecology Letters*, 10(9):E4–E8, 2007.
38. S. Letovsky. Cognitive processes in program comprehension. In *Empirical Studies of Programmers*, pages 58–79, 1986.
39. Che Ma and Geoffrey Chang. Retraction for Ma and Chang, Structure of the multidrug resistance efflux transporter EmrE from *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 104(9):3668, 2007.
40. Robert C. Martin. *Agile Software Development, Principles, Patterns, and Practices*. Prentice Hall, 2002.
41. David Matthews, Greg Wilson, and Steve Easterbrook. *Configuration Management for Large-Scale Scientific Computing at the UK Met Office*. *Computing in Science & Engineering*, November-December 2008.
42. Steve McConnell. *Code Complete: A Practical Handbook of Software Construction*. Microsoft Press, 2 edition, 2004.
43. Zeeya Merali. Error: Why Scientific Programming Does Not Compute. *Nature*, 467:775–777, 2010.
44. Gerard Meszaros. *xUnit Test Patterns: Refactoring Test Code*. Addison-Wesley, 2007.
45. William Stafford Noble. *A Quick Guide to Organizing Computational Biology Projects*. *PLoS Computational Biology*, 5(7), 2009.
46. William L. Oberkampf and Christopher J. Roy. *Verification and Validation in Scientific Computing*. Cambridge University Press, 2010.
47. Open Provenance. <http://openprovenance.org>. Viewed June 2012.
48. Andy Oram and Greg Wilson, editors. *Making Software: What Really Works, and Why We Believe It*. O'Reilly, 2010.
49. Roy Osherove. *The Art of Unit Testing: With Examples in .NET*. Manning, 2009.
50. Joe Pitt-Francis, Miguel O. Bernabeu, Jonathan Cooper, Alan Garny, Lee Momtahan, James Osborne, Pras Pathmanathan, Blanca Rodriguez, Jonathan P. Whiteley, and David J. Gavaghan. Chaste: Using Agile Programming Techniques to Develop Computational Biology Software. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1878):3111–3136, September 2008.
51. Yann Pouillon, Jean-Michel Beuken, Thierry Deutsch, Marc Torrent, and Xavier Gonze. Organizing Software Growth and Distributed Development: The Case of Abinit. *Computing in Science & Engineering*, 13(1):62–69, 2011.
52. Prakash Prabhu, Thomas B. Jablin, Arun Raman, Yun Zhang, Jialu Huang, Hanjun Kim, Nick P. Johnson, Feng Liu, Soumyadeep Ghosh, Stephen Beard, Taewook Oh, Matthew Zoufaly, David Walker, and David I. August. A Survey of the Practice of Computational Science. In *24th ACM/IEEE Conference on High Performance Computing, Networking, Storage and Analysis*, 2011.
53. Lutz Prechelt. Two Comparisons of Programming Languages. In Andy Oram and Greg Wilson, editors, *Making Software: What Really Works, and Why We Believe It*. O'Reilly, 2010.
54. Deborah S. Ray and Eric J. Ray. *Unix and Linux: Visual QuickStart Guide*. Peachpit Press, 4 edition, 2009.
55. Evan Robinson. Why Crunch Mode Doesn't Work: Six Lessons. <http://www.igda.org/why-crunch-modes-doesnt-work-six-lessons>, 2005. Viewed June 2012.
56. R. Sanders and D. Kelly. Dealing with Risk in Scientific Software Development. *IEEE Software*, 25(4):21–28, July-August 2008.
57. J. Segal. Models of Scientific Software Development. In *First International Workshop on Software Engineering for Computational Science and Engineering*, 2008.
58. J. Segal and C. Morris. Developing Scientific Software. *IEEE Software*, 25(4):18–20, 2008.
59. Judith Segal. When Software Engineers Met Research Scientists: A Case Study. *Empirical Software Engineering*, 10(4):517–536, 2005.
60. Peter Smith. *Software Build Systems: Principles and Experience*. Addison-Wesley, 2011.
61. Joel Spolsky. The Joel Test: 12 Steps to Better Code. <http://www.joelonsoftware.com/articles/fog0000000043.html>, 2000. Viewed June 2012.
62. Burak Turhan, Lucas Layman, Madeline Diep, Hakan Erdogmus, and Forrest Shull. How Effective is Test-Driven Development? In Andy Oram and Greg Wilson, editors, *Making Software: What Really Works, and Why We Believe It*, pages 207–217. O'Reilly, 2010.
63. Moshe Vardi. Science Has Only Two Legs. *Communications of the ACM*, 53(9), September 2010.
64. Laurie Williams. Pair Programming. In Andy Oram and Greg Wilson, editors, *Making Software: What Really Works, and Why We Believe It*, pages 311–322. O'Reilly, 2010.
65. Greg Wilson. Software Carpentry: Getting Scientists to Write Better Code by Making Them More Productive. *Computing in Science & Engineering*, November-December 2006.
66. List of unit testing frameworks. http://en.wikipedia.org/wiki/List_of_unit_testing_frameworks. Viewed June 2012.
67. Andreas Zeller. *Why Programs Fail: A Guide to Systematic Debugging*. Morgan Kaufmann, 2009.