John Jörholt

Assignment 1

The main point that all the authors of the five papers agree on is that you should do your research before starting to develop a new program. Most likely someone has done the same or similar thing that you are looking to do. They also talk about the importance and benefits of collaboration and open source development, you should share your code openly, take inspiration from others, and encourage others to help and improve your code.

Most of the papers also talk about version control and its perks both for organization and for making it easy to correct mistakes. With version control you can track which changes you did to your code and when you did them. It also allows you to restore previous versions of your code if you made a change to your code that changed the way it work or produces unexpected errors.

Paper 1 (Noble, 2009)

The first paper brings up the importance of organizing your files and directories so both you and your colleges can easily go back and check what experiments were done, when they were done, and which code and protocols were used. It also recommends using a lab notebook organized chronologically in combination with the organized file system. Keeping a notebook with comments and useful notes will help you greatly when you need to go back and check on things you did months or maybe years earlier.

The paper also talks about the differences between command line commands, scripts and programs and when different approaches are most suitable.

Paper 2 (Sandve et al., 2013)

The second paper brings up the importance of keeping track of your data and how your results were produced. It recommends documenting how every result were produced, avoiding steps of manual data manipulation, archiving the exact versions of external programs that have been used, record all intermediate results, note the underlying seeds when using randomness in your analysis, always store the raw data behind your plots, generate hierarchical analysis output, and connecting textual statements to underlying results.

Paper 3 (Loman & Watson, 2013)

The third paper recommends testing your code using small datasets with known answers so you can check that your code produces the correct result for all kinds of inputs. It also mentions that you should remember that you are a scientist and not a programmer; your code does not have to look pretty as long as it works. Focus on creating code that works first and then make it prettier by commenting and documenting how it works.

The authors also mention that you should be suspicious of interesting results, your data will always produce both false positives and negatives so test your data using multiple approaches.

Paper 4 (Wilson et al., 2014)

The fourth paper talks about writing programs for people not computers, meaning that you should make your code a readable as possible. The readers should not have to hold more than a handful of facts in memory, make names consistent, distinctive, and meaningful, and keep the style and format of your code consistent. It also recommends that you make incremental changes and that you should plan for mistakes.

Paper 5 (Prlic & Procter, 2012)

The fifth paper advises you to keep it simple, don't make your software to complicated for others to install and use. The authors also tells you to not be a perfectionist, share your code early and often with others, this will help you identify bugs and will help you improve your code faster. They also bring up the importance of promoting your project as often and as much as possible, both online and on conferences our similar venues, and finding sponsors through grants for example. Lastly they remind you that the development of software should primarily be a way for you to advance your research.

John Jörholt

Assignment 2

ANNOVAR (annovar.openbioinformatics.org/en/lastest/) is a software tool that utilizes update-to-date information to functionally annotate genetic variants detected from diverse genomes. It can perform gene-based, region-based, and filter-based annotation, as well as a number of other functionalities. ANNOVAR is written in Perl and the program package contains the main program annotate_variation.pl together with five accessory programs.

When no parameters are supplied to the main program annotate_variation.pl it prints a usage description with description of arguments and examples of how to use the different functions of the program. It has a –help switch but it does not have a –version switch, however the version is printed on the bottom of the usage description when no parameters are given to the program.

ANNOVAR uses stderr for error messages and prints error messages when something goes wrong or if there is something wrong or missing in the parameters.

As I can tell the program does not contain any hardcoded paths and does not seem to pollute the PATH. The program does not use any external tools but it does rely on downloaded databases, the program checks and throws an error if a database in not found.

In my opinion ANNOVAR is a very useful and user-friendly tool for annotating your genetic variation data. It has a easy learning curve and tells you in an good way if you do something wrong. It is also updated frequently and the supported databases are expanded often.