Week 5, Lecture 9

István Albert

Biochemistry and Molecular Biology
and Bioinformatics Consulting Center
Penn State

Quality Control and Filtering

- Removing or altering the data based on objective measures

- Isn’t that data massaging/altering?

- Good question: NO – but needs to be done solely using instrument produced (analysis independent) factors

Understanding the FastQ format

Also available on the FastQ wikipedia page originally from
For the MAC: homebrew

It allows you to install some libraries and tools that will be required later. Linux already has package managers `apt-get, yum` etc.

Steps to installing tools

Determine the distribution type

1. Executable (static binary) code. Download the code and you are done.

2. Source code. Download the code and go to next step.

Determine the type of the source code

1. Source is for a **compiled language** that will be turned into a binary program (typically C but could be others)

2. Source is for an **interpreted language** that will run the code: java, perl, python, ruby

Check list for source code that needs compilation

1. Does it have a `configure` script? If yes then run it
   ```
   ./configure
   ```

2. Run `make`
   ```
   make
   ```

Ideally you should be done. This will create the binary. The program may need library dependencies. If so those
Checklist for interpreted languages

1. You need to have the language installed. Most modern computers have perl, python, java installed by default.

2. The source code may have “dependencies” – a much dreaded word could lead to a lengthy procedure of downloading other code that in turn may depend on other and other etc...

Automated installation

- Language specific – will require installing a language specific package manager
- Python has pip, Perl has MCPAN, ruby has gem
- Using these is outside of the scope of these lectures. It would look like this:

  `pip install package-name`

Installing good tools is very easy – not so good ones are mini puzzles – badly designed tools are incredibly frustrating

Back to quality control

- Modify the fastq records to remove data that was labeled as being inaccurate

  Typical operations are to

  - remove reads
  - shorten reads

Fastq Quality Shootout
Fastx toolkit contains a large number of tools:

- Tools that run on both Fasta and Fastq formats have a name that starts with fastx.
- Tools that only run on Fastq format have a name that starts with fastq.

FastX works by default with the phred 64 encoding.

The FastQ files can have different encodings. An older (2 year) instrument may work with a different encoding and will probably be around for a decade.

Most tools will not detect this and will either at best crash mysteriously or just work happily and at worst produce wrong results.
Other FastX tools

This tool removes duplicated reads.
About 40% of data!
But also turns the file into FASTA.

Tool List

- **Seqtk** – binary use: make
- **NGS Toolkit** – perl, has good manual
- **Trimmomatic** – java, somewhat obscure usage
- **Prinseq** – beatiful manual and website, more than 20 times slower than the others
- **Biopieces** – its is not a tool it is more of a lifestyle. Lots of installation steps.

Other tools

- See the **Shootout** – serves as supplementary information
- Quality control often goes way beyond read manipulation and can be thought as a pre-analysis – at that point it should not be called QC though.
- Some tools may have particular features that directly apply to your research.

Homework 9

- Enumerate three tools in the Fastx toolkit
- Using one of the tools presented in the lecture try to make one of the datasets a better
- Describe the process that you have applied and at least one observable change in the data.
- Include an amazing “before” and “after” quality screenshot.