2012 - BMMB 597D: Analyzing Next Generation Sequencing Data

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

<table>
<thead>
<tr>
<th>33</th>
<th>59</th>
<th>64</th>
<th>73</th>
<th>104</th>
<th>126</th>
</tr>
</thead>
</table>

*S* - Sanger  Phred+33, raw reads typically (0, 40)

*X* - Solexa  Solexa+64, raw reads typically (-5, 40)

*I* - Illumina 1.3+ Phred+64, raw reads typically (0, 40)

*J* - Illumina 1.5+ Phred+64, raw reads typically (3, 40)  
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)  
(Note: See discussion above).

*L* - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

Also available on the FastQ wikipedia page originally from

More on FastQC interpretation
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
Tutorial: Fastq Quality Control Shootout

This tutorial also serves as the supporting information for Lecture 9 of the course titled Analyzing High Throughput Sequencing Data offered at Penn State.

Within this tutorial we compare the efficiency and characteristics of several software tools designed to filter FASTQ files. We choose two simple but common operations and perform them with different tools. For some tasks we will provide what we call a naive implementation in various languages. These are implementations that would take a programmer about 5-10 minutes to create. These are for illustrative purposes only.

We want to make a note that the Fastx Toolkit exhibits a strange behavior in that the binary code downloaded for a Mac runs more than five times slower than the binary code compiled locally on the same computer. The times are displayed for the faster version of the Fastx Toolkit. For the NGS Toolkit we have disabled the (otherwise quite helpful) FastQ format detection subroutines.

The test data `s1.fq` contains 1 million FASTQ records.

All times are reported in seconds (s). Tools are ordered by runtime.

### Clipping Sequences

Clipping is removing parts of each fastq record. For this test we’ll remove 10 bases from the start and end.

2.5s with Seqtk:

```bash
seqtk trimfq -b 10 -e 10 data/s1.fq > data/tmp.fq
```

3.7s with Fastx Toolkit
Fastx Toolkit

One of the first binary quality filtering packages

Here you’ll find a short description and examples of how to use the FASTX-toolkit from the command line.

- **Command Line Arguments**
  - FASTQ-to-FASTA
  - FASTQ/A Quality Statistics
  - FASTQ Quality chart
  - FASTQ Quality chart
  - FASTQ/A Nucleotide Distribution chart
  - FASTQ/A Clipper
  - FASTQ/A Renamer
  - FASTQ/A Trimmer
  - FASTQ/A Collapser
  - FASTQ/A Artifacts Filter
  - FASTQ Quality Filter
  - FASTQ/A Reverse Complement
  - FASTQ Formatter
  - FASTA nucletides changer
Composed of multiple scripts


Part of FASTX Toolkit 0.0.13 by A. Gordon (gordon@cshl.edu)

- **[-h]** = This helpful help screen.
- **[-q N]** = Minimum quality score to keep.
- **[-p N]** = Minimum percent of bases that must have [-q] quality.
- **[-z]** = Compress output with GZIP.
- **[-i INFILE]** = FASTA/Q input file. default is STDIN.
- **[-o OUTFILE]** = FASTA/Q output file. default is STDOUT.
- **[-v]** = Verbose - report number of sequences.
  If [-o] is specified, report will be printed to STDOUT.
  If [-o] is not specified (and output goes to STDOUT), report will be printed to STDERR.
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
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
Tool List

• **Seqtk** – binary use: make
• **NGS Tookit** – 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.
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.