A bioinformatician’s job requires them to evaluate and install a large number of tools. The ease of installation usually correlates with the quality of tool. Documentation is essential; otherwise, it is no more than a black box.

Package manager 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 (binary) code.**
   Download the code and you are done. Easy to install → may not be optimized to your system

2. **Source (text) code.**
   Download the code and **compile it** (see next slides)
**Determine the type of the source code**

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

2. Source is of 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. Now run make
   
   make

   Ideally you should be done. This will create the binary.

   (The program may need library dependencies. Then those need to be installed as above)

**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 easy_install and pip, Perl has MCPAN, ruby has gem

   ```
   easy_install install package-name
   or
   pip install package-name
   ```

   Installing good tools is very easy – not so good ones are mini puzzles – badly designed tools are incredibly frustrating.
Quality Control and Filtering

- Removing or altering the data based on objective measures
- Isn’t that data massaging?
- Good question – one needs to be very careful not to bias data

Understanding sequencing

- Library prep has many steps
- Sequencing may introduce artifacts
- Always try to understand what the instrument does and may happen when things are not optimal
- See Short Guide to Illumina sequencing on the webpage

Random DNA fragment sequencing with Illumina

- Fragmentation
- Adapter Ligation
- Sequencer
- One read

FastQC report shows biases

- Per base sequence quality
- Base Content

Our job is to fix this and we need to install tools for that
Quality control operations

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

Typical operations are to

- remove (discard) reads - **careful with this**!
- shorten reads (trimming) by quality or by removing patterns

Tool List

- **Seqtk** – fastest tool
- **Cutadapt** – adaptor cutting
- **NGS Toolkit** – perl, has good manual
- **Trimmomatic** – java, somewhat obscure usage
- **Prinseq** – beatiful manual and website, appears to be slow
- **Biopieces** – its is not a tool it is more of a lifestyle. Lots of installation steps.

Fastq Quality Shootout

Tool List

- **Seqtk** – fastest tool
- **Cutadapt** – adaptor cutting
- **NGS Toolkit** – perl, has good manual
- **Trimmomatic** – java, somewhat obscure usage
- **Prinseq** – beatiful manual and website, appears to be slow
- **Biopieces** – its is not a tool it is more of a lifestyle. Lots of installation steps.

Installation

One time tasks (see code repository):

- Mac: install **Homebrew**
- Using **Homebrew** install **git**
- Using **easy_install** install **pip**

Use **git** and **pip** to install tools in the future

- clone the **seqtk** repository with **git** and **make** it, link to ~/bin
- **pip** install cutadapt, does not need to be linked
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 11

- Install **cutadapt** and **seqtk**

- Use data **sample1.fq** and **sample2.fq** distributed with **Lecture 10**

- Remove adapters with **cutadapt** and/or trim your sequences with **seqtk** (you may use other tools as well)

- Run **FASTQC** on the cut/trimmed data. Select a plot from each report and explain the differences that you see.