Sequence QA and Cleaning

BCB 504: Applied Bioinformatics

Matt Settles

University of Idaho
Bioinformatics and Computational Biology Program

January 29, 2015

Outline

QA

Its best to perform QA on both the run as a whole (poor samples can affect other samples) and on the samples themselves. Reports such as Basespace for Illumina, are great ways to evaluate the runs as a whole.

QA on the sample data can occur using 3rd party applications that evaluate quality.
Such as the fastqc application.

Why Clean Reads

We have found that aggressively "cleaning" and processing reads make a large difference to speed and quality of assembly and mapping results. Cleaning your reads means, removing reads/bases

that are:

- not of primary interest (contamination)
- originate from PCR duplication
- artificially added onto sequence of primary interest (vectors, adapters, primers)
- low quality bases
- other unwanted sequence (polyA tails in RNA-seq data)
- join short overlapping paired-end reads

Cleaning reads, some strategies I

- **Quality trim/cut**
  - 'end' trim a read until the average quality $>$ Q (Lucy)
  - remove any read with average quality $<$ Q
- **eliminate singletons**
  - If you have excess depth of coverage, and particularly if you have at least x-fold coverage where x is the read length, then eliminating singletons is a nice way of dramatically reducing the number of error-prone reads.
- **eliminate all reads (pairs) containing an N**
  - If you can afford the loss of coverage, you might throw away all reads containing Ns.
- **Identity and trim off adapter and barcodes if present**
  - Believe it or not, the software provided by Roche or Illumina, either does not look for, or does a mediocre job of, identifying adapter and removing them.

Cleaning reads, some strategies II

- **Identity and remove contaminant and vector reads**
  - Reads which appear to fully come from extraneous sequence should be removed.

preproc_experiment

preproc_experiment is a command line pipeline (Written in R) that employs most of the techniques described above. It can be found on the IBEST CRC servers in the grc/2.0 module.

Aggressive cleaning and filtering of raw reads

Its better to have low coverage of really good quality sequence data, than high coverage of poor quality data

1. Remove PCR duplicates (we use bases 10-35 of each read)
2. Remove/Trim Contaminants (at least PhiX) and Vectors (if used)
3. Search for and remove Illumina adapters
4. Trim sequences (Left and Right) by quality score (I like Q24)
5. If RNA and if mapping, trim polyA/T
6. Join and extend, overlapping paired end reads
7. Filter out rRNA genes (if ribosomal depletion was performed)

Cleaning as QA

Beyond generating better data for downstream analysis, cleaning statistics also give you an idea as to the quality of the sample, library generation, and sequencing technique used to generate the data. It can help inform you of what you might do in the future.