Sequence QA and Cleaning

BCB 504: Applied Bioinformatics

Matt Settles

University of Idaho
Bioinformatics and Computational Biology Program

February 25, 2013

Outline

Yeast dataset and folder structures

The Yeast dataset we will use for all project is located on the CRC servers:
/mnt/home/uirig/user_data/BCB504 There are 4 Roche 454 runs

that include Yeast sample (see 454_runs_sample_sheet for details) and 2 Illumina MiSeq runs that include Yeast Sample.

The Yeast we sequenced was from Jill Johnson's lab (UI) and is a mutant of the W303 strain.

QA

Its best to perform QA on both the run as a whole (poor samples [barcodes] can affect other samples) and on the samples themselves. Reports (Basespace for Illumina) and Roche QA data in the SignalProcessing folder are great ways to evaluate the runs as a whole. QA on the sample data can occur using 3rd party

applications that evaluate quality.
for Roche, I have an R script that produces reports from Raw Sff files, available:
/mnt/home/uirig/roche454/runQA-metrics.R
For Illumina, we currently use the fastqc application.

Why Clean Reads

While it does not seem to be reported anywhere, we have found that "cleaning" your reads make a large difference to speed and quality of assembly and we suspect mapping results too. Cleaning

your reads means, removing bases that are:

- not of primary interest (contamination)
- artificially added onto sequence of primary interest (vectors and adapters)
- low quality bases
- other unwanted sequence (polyA tails in RNA-seq data)

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**

Cleaning reads, some strategies II

- Believe it or not, the software provided by Roche or Illumina, either does not look for, or does a mediocre job of, identifing adapter and removing them.

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

seqyclean

Seqyclean is an application created by Ilya (in class) that employs most of the techniques described above. It can be found:

seqyclean

and is installed on the IBEST CRC servers in the grc module.

```
====================Summary Statistics====================
Reads analyzed: 1068653, Bases:521817306
Found ->
Left mid tag: 1068653, 100%
Right mid tag: 1067064, 99.8513%
# of reads with vector: 1065914, 99.7437%
Reads left trimmed ->
By adapter: 20962
By quality: 7465
By vector: 323056
Average left trim length: 29.3128 bp
Reads right trimmed ->
By adapter: 115195
By quality: 234603
By vector: 924
Average right trim length: 175.871 bp
Reads discarded: 740919 ->
By read length: 740919
----------------------------------------------------------
Reads accepted: 327734, %30.668
Average trimmed length: 325.645 bp
==========================================================
Program finished.
Elapsed time = 1.512389e+03 seconds
```

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.