FASTQ Manipulation Tools 2014

Trimming low quality bases

There are a number of open source tools that can trim off 3' bases and produce a FASTQ file of the trimmed reads to use as input to the alignment program.

FASTX Toolkit

The **FASTX-Toolkit** provides a set of command line tools for manipulating **fasta** and **fastq** files. The available modules are described on their website. They include a fast **fastx_trimmer** utility for trimming fastq sequences (and quality score strings) before alignment.

FASTX-Toolkit is available via the TACC module system.

FASTX_toolkit module description

```
module spider fastx
module load fastx_toolkit
```

Let's run **fastx_trimmer** to trim all input sequences down to 90 bases:

Run fastx_Trimmer

```
fastx_trimmer -i data/Sample1_R1.fastq -l 90 -Q 33 -o Sample1_R1.trimmed.fastq
```

- The **-l 90** option says that base 90 should be the last base (i.e., trim down to 50 bases)
- the **-Q 33** option specifies how base qualities on the 4th line of each fastq entry are encoded. The FASTX toolkit is an older program, written in the time when Illumina base qualities were encoded differently. These days Illumina base qualities follow the Sanger FASTQ standard (Phred score + 33 to make an ASCII character).

Exercise: fastx toolkit programs

What other fastx manipulation programs are part of the fastx toolkit?

Type **fastx**_ then tab to see their names
See all the programs like this:

fastx toolkit programs

```
ls $TACC_FASTX_BIN
```

Exercise: What if you just want to get rid of reads that are too low in quality?

fastx_quality_filter syntax

```
fastq_quality_filter -q <N> -p <N> -i <inputfile> -o <outputfile>
-q N: Minimum Base quality score
-p N: Minimum percent of bases that must have [-q] quality
```

Let's try it on our data- trim it to only include reads with atleast 80% of the read having a quality score of 30 or above.

Run fastx_quality_filter

```
fastq_quality_filter -q 20 -p 80 -i data/Sample1_R1.fastq -Q 33 -o Sample1_R1.filtered.fastq
```

Exercise: Compare the results of fastq_trimmer vs fastq_quality_filter

Compare results

```
grep '^@HWI' Sample1_R1.trimmed.fastq |wc -l
grep '^@HWI' Sample1_R1.filtered.fastq |wc -l
```

Adaptor Trimming

Data from RNA-seq or other library prep methods that resulted in very short fragments can cause problems with moderately long (50-100bp) reads since the 3' end of sequence can be read through to the 3' adapter at a variable position. This 3' adapter contamination can cause the "reql" insert sequence not to align because the adapter sequence does not correspond to the bases at the 3' end of the reference genome sequence.

Unlike general fixed-length trimming (e.g. trimming 100 bp sequences to 40 or 50 bp), adapter trimming removes differing numbers of 3' bases depending on where the adapter sequence is found.

The GSAF website describes the flavaors of Illumina adapter and barcode sequence in more detail https://wikis.utexas.edu/display/GSAF/Illumina+-+all+flavors

FASTX Toolkit

One of the programs available as part of the fastx toolkit does a crude job of clipping adaptors out of sequences.

fastx_clipper will clip a certain nucl. sequence (eq: adapter) from your reads.

fastx_Clipper general syntax

```
fastx_clipper -a <adapter> -i <inputfile> -o <outputfile> -l <discardSeqsShorterThanN>
```

Cutadapt

The **cutadapt** program is an excellent tool for removing adapter contamination. The program is not available through TACC's module system but we've installed a copy in our $BI/bin directory. Cutadapt has some advantages over fastx_clipper:

- Cutadapt allows for mismatches.
- Cutadapt allows for paired end support.

cutadapt general syntax

```
cutadapt -a <adapter> -e <errorRate> -m <minLength> -o <outputFile> <InputFile>
```

When you run **cutadapt** you give it the adapter sequence to trim, and this is different for R1 and R2 reads.

cutadapt command for R1 sequences

```
/corral-repl/utexas/BioITeam/bin/cutadapt-1.3/bin/cutadapt -m 22 -a GATCGGAAGAGCACACGTCTGAACTCCAGTCAC
Sample1_R1.fastq
```

cutadapt command for R2 sequences

```
/corral-repl/utexas/BioITeam/bin/cutadapt-1.3/bin/cutadapt -m 22 -a TGATCGTCGGACTGTAGAACTCTGAACGTGTAGA
Sample1_R1.fastq
```

Notes:

- The **-m 22** option says to discard any sequence that is smaller than 22 bases after trimming. This avoids problems trying to map very short, highly ambiguous sequences.

Paired end commands are a little more complicated: It's a multi step process: first providing both R1 and R2 files to create tmp files which are input to the second cutadapt command.

cutadapt command for R2 sequences

```
/corral-repl/utexas/BioITeam/bin/cutadapt-1.3/bin/cutadapt -a GATCGGAAGAGCACACGTCTGAACTCCA -m 22 --paired-
output tmp.2.fastq -o tmp.1.fastq  Sample1_R1.fastq Sample1_R2.fastq &
/corral-repl/utexas/BioITeam/bin/cutadapt-1.3/bin/cutadapt -a ATCGTCGGACTGTAGAACTCTGAACGTG -m 22 --paired-
output trimmed.1.fastq -o trimmed.2.fastq tmp.2.fastq tmp.1.fastq &
rm tmp.1.fastq tmp.2.fastq
```

Please refer to https://wikis.utexas.edu/display/GSAF/Illumina+-+all+flavors for Illumina library adapter layout.

Appendix: Illumina Adapter Information

https://wikis.utexas.edu/display/GSAF/Illumina+-+all+flavors

<P5 primer/capture site> <IndexRead2> <Read1 primer site>
<template - gDNA, RNA, amplicon, whatever>
<Read2 primer site> <IndexRead1> <P7 primer/capture site>

Standard DNA Library

- Read 1- Look for <Read 2 primer site>

GATCGGAAGAGCACACGTCTGAACTCCAGTCAC

- Read 2 - Look for <RevComp of TruSeq Read 1 primer>

GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGA

Small-rna or RNA library

- Read 1- Look for <Read 2 primer site>

GATCGGAAGAGCACACGTCTGAACTCCAGTCAC

- Read 2 - Look for <RevComp of Read 1 primer site(NEB)>

TGATCGTCGGACTGTAGAACTCTGAACGTGTAGA

BACK TO COURSE OUTLINE