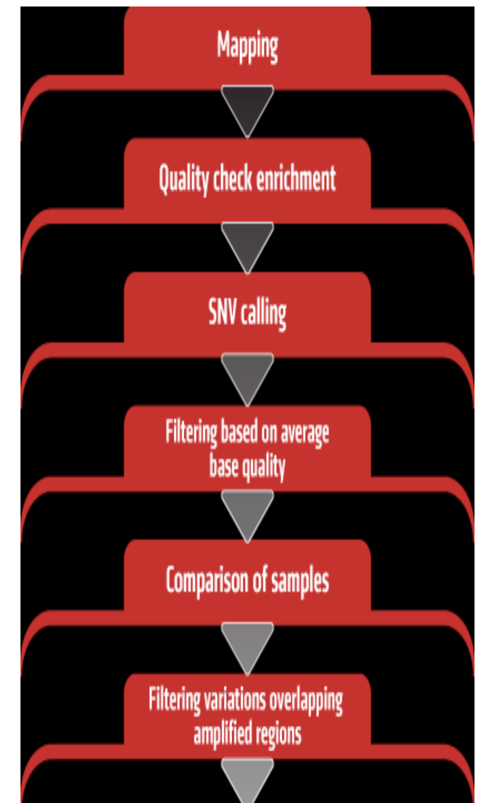
A Quick Introduction to Fastx Toolkit and Cutadapt

Dhivya Arasappan
Byte Club
11-12-13

THE UNIVERSITY OF
TEXAS
AT AUSTIN

- In a typical NGS data analysis workflow:
  - Check quality using FastQC
  - **Preprocess fastq/fasta files**
    - Fastx toolkit
    - Cutadapt



- Fastx toolkit Available on:
  - Fourierseq
  - Lonestar - module load fastx_toolkit

- Cutadapt available on Fourierseq, needs to be installed on lonestar.

More info on Fastx toolkit: http://hannonlab.cshl.edu/fastx_toolkit

Few Useful Commands

If your data has significant adapter contamination

- fastx_clipper
  - To clip a certain nucl. Sequence (eg: adapter) from your reads

>Read1
TTGGTATCCCAACAGCGTACCCGCC
GGTAAAGGGTCCAAGAACTCACACT
TCGACAAGACCTCGAGA**GATCGGAA**
GAGCACACGTCTGAACTCCA

>Read1
TTGGTATCCCAACAGCGTACCCGCC
GGTAAAGGGTCCAAGAACTCACACT
TCGACAAGACCTCGAGA

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

An alternative to fastx_clipper for handling adapter contamination

- Cutadapt
  - Allows for mismatches
  - Allows for paired-end support

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

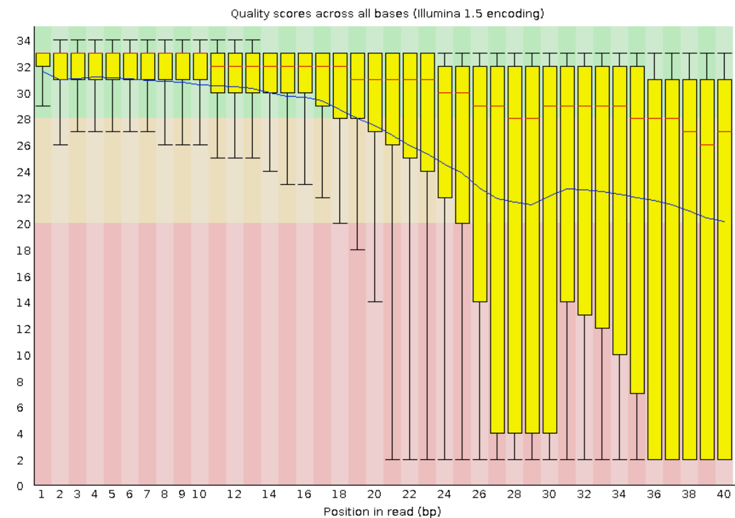
Few Useful Commands

If your data has an artifact at a certain position

- fastx_trimmer
  - To trim bases from the beginning or end of all your reads

fastx_trimmer -i <inputfile> -o <outputfile> -f <FirstBaseToKeep> -l <LastBaseToKeep>

If your data has low quality reads



- fastq_quality_filter
  - To filter low quality reads

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.

Things to Remember and Other Modules

If your data is generated by GSAF (or using Illumina pipeline 1.8 and higher)

- Fastq file will be in Sanger format
- Add *-Q33* flag when running fastx toolkit modules

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                            104                  126
0........................26...31.......40
                         -5....0........9.............................40
                               0........9.............................40
                                  3.....9.............................40
0.2......................26...31........41

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)
```

Other things you can do…

- Convert fastq to fasta
- Format multiline fasta file
- Generate nucleotide distribution

```
>Read1
ACCATACTNGA
CATTAGGACAT
GACCG
```

```
>Read1
ACCATACTNGCATTAGGACATGACCG
```

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