Bisonnet Help Page

Chase Hoehn

Program: seqtk

Module: seq_tools

Version: 1.3

What it Does:

This program is a tool used for processing FASTA and FASTQ files. One of the most common uses of this program is to convert a FASTQ file into a FASTA file. The difference between these file formats are as follows:

FASTQ:

The FASTQ format is a format that contains raw sequence reads in addition to base quality scores (which tell us the confidence/likelihood that the nucleotide the sequence detects is the true nucleotide of that sequence). This format is used in the output files of many different sequencers, but is usually seen via the output of Illumina sequencer files. The format was created as a way to note the confidence in each base call/nucleotide (due to different sequencing technologies having varying confidence levels). The start of a new sequence is denoted by "@," and the start of the quality scores of the associated sequence is denoted by "+" following the sequence. An example of a FASTQ format file is shown below.

```
[cfh007@bisonnet-hpc Practical_3]$ head Ppyra1_antenna_500k_1.fq
@A01050:44:HWNGNDSXX:2:1605:8992:30718 1:N:0:TGGAAT
CACTTTCTACAATGAGCTGCGTGTCGCACCAGAAGAACACCCCATCCTTCTCACCGAAGCTCCACTCAACCCCAAGGCTAACCGTGAAAAGATGACCCAAATCATGTTTGAAACCTTCAACACCCCAGCCATGTACGTCGCCATTCAGGC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A01050:44:HWNGNDSXX:2:2320:2926:33426 1:N:0:TGGAAT
GTTTTTTACTAATGCTTCATTAACTAGTTTCTGCCAATTCTTAAAATACTTTGGGCCATGCCCATATGCATTGCGAATACTAGCTTCGGCATTTACTCTCGTTGCTTCTTTTGCATCATGTGATTGCCAAATCCACTGAGCACTTTCGTG
+
FFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFF:FFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFF:FFFF:FFFFFFF:F:FFF
```

FASTQ Quick Identifier:

Line 1: This begins with the "@" character followed by a sequence identifier and sometimes a description.

Line 2: This is the raw sequence letters.

Line 3: This line starts with the "+" character which separates the sequence from the following line.

Line 4: This line contains the associated quality scores for each nucleotide base call from line 2.

FASTA:

The FASTA format is a different format that contains raw sequence reads of DNA or protein sequences. This format does not contain quality scores for each nucleotide. This format starts with the ">" symbol on the first line which is followed by the sequence ID. The second line denotes the start of the sequence. An example of a FASTA format file is shown below.

```
[[cfh007@bisonnet-hpc Practical_7]$ head Ecorr_transcriptome.fasta
>evm-Ecorr_v1_Scaffold100425-processed-gene-0.0-mRNA-1  [seqid Ecorr_v1_Scaffold100425:926-1754 +]
ACAGTCGTGGAGGTTTGCCAAAAGAAGCGACAATGCAAATTTAAGTCTTCGCCGAGAACG
TTTGGGGGAAATCCGTGCCCGGGCGTTCGTAAATATGTGGAAGTTGCCTATAAGTGCAGG
CCATACGAATTTCGAAGCAAGGTGGCATGCCAAAACGAAGGAATACAACTAAGTTGCAAT
CCGAATTCGAGAATTGCGATTTATAGCGCCAGTTATGGAAGGACAGAGTATGAAAGCATA
CAATGTCCTCAACCTCAAGGAGTTCCGGAAGAAACTTGTTTGGTAAGTTATGGTACCGAG
ACGGTGATGAAGCTCTGCCATGGGAAAAGAACTTGTGATGTTTTGGCAGACATTAAAACT
TTCGGCAGTCCGTGTCGCCCAGAATCGCGGATGTACTTAAAAGTTATTTACACCTGCGTC
CCACGGAAGGTGTTAAAGGAGCAGTTCGAAGGCCAACCTGAACCAGATGAGCTGACTAAC
GAAATGGAAAATTCCGACATCGACGACGATTTTGAAAATTATGATGCTGGGGATGAGTTT
```

FASTA Quick Identifier:

Line 1: This begins with the ">" character which is followed by a sequence identifier and often a description of the sequence.

Line 2: This is the start of the raw sequence letters/nucleotide base calls.

How to Use:

Note: This is a "lightweight" program that runs very quickly. Thus, it can be run interactively (i.e. it will run almost instantaneously and doesn't require a script to be submitted to the queue).

Note: *Italicized* text denotes sample code for each step.

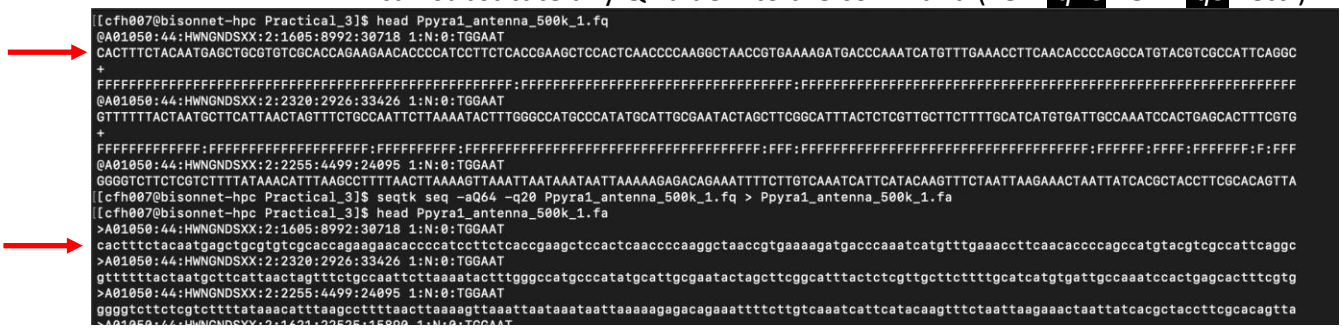
1. Log into Bisonnet, then connect to a compute node
    a. *ssh cfh007@bisonnet-hpc.bucknell.edu*
    b. *srun -n 1 -p short --pty /bin/bash*
2. Load the module
    a. *module load seq_tools*
3. Locate a FASTQ format file and navigate to its directory
    a. File name format looks like "____.fastq" OR "___.fq"
4. Convert FASTQ format file into a FASTA format file
    a. *seqtk seq -a input.fq > output.fa*
    b. Example code below. ("Ppyra1_antenna_500k_1.fq" is converted into "Ppyra1_antenna_500k_1.fa")

```
[[cfh007@bisonnet-hpc Practical_3]$ ls
fastqc_output                 Ppyra1_antenna_500k_1_unpaired.fq  Ppyra1_antenna_500k_2_unpaired.fq  trimmomatic.sh
Ppyra1_antenna_500k_1.fq      Ppyra1_antenna_500k_2.fq           slurm.hpc-1.2389.stderr.txt        trinity_output
Ppyra1_antenna_500k_1_paired.fq  Ppyra1_antenna_500k_2_paired.fq  slurm.hpc-1.2389.stdout.txt        trinity.sh
[[cfh007@bisonnet-hpc Practical_3]$ module load seq_tools
NOTE: Module requires Python 3.7. All other Python versions have been unloaded.

Loading seq_tools/2020.01.13
  Loading requirement: cuda10.0/toolkit/10.0.130 python/3.7 perl/5.30.1
[[cfh007@bisonnet-hpc Practical_3]$ seqtk seq -a Ppyra1_antenna_500k_1.fq > Ppyra1_antenna_500k_1.fa
[[cfh007@bisonnet-hpc Practical_3]$ ls
fastqc_output                 Ppyra1_antenna_500k_1_paired.fq    Ppyra1_antenna_500k_2_paired.fq    slurm.hpc-1.2389.stdout.txt  trinity.sh
Ppyra1_antenna_500k_1.fa      Ppyra1_antenna_500k_1_unpaired.fq  Ppyra1_antenna_500k_2_unpaired.fq  trimmomatic.sh
Ppyra1_antenna_500k_1.fq      Ppyra1_antenna_500k_2.fq           slurm.hpc-1.2389.stderr.txt        trinity_output
```

Other helpful features:

- Convert FASTQ format file into a FASTA format file while also turning any base calls with a Q-score (quality score) below 20 (or any other Q-score) into lowercase
    - *seqtk seq -aQ64 -q20 input.fq > output.fa*
    - Example code below. **Note**: all bases are lowercase because quality scores are mostly "F" which is equal to Q=6 (which is less than Q=20). You can substitute any Q value into the command (i.e. *"-q10"* OR *"-q5"* etc.)



- Make the reverse compliment of sequence reads for FASTA or FASTQ format files
    - *seqtk seq -r input.fq > output.fq*
    - Sequence compliments are shown running in reverse directions