Working with FASTQ and FASTA data

Edit | New Page

Jump to bottom

AmyStonelake edited this page 2 hours ago · 21 revisions

Always remember to activate the bioinfo environment when working on Biostar class materials.

```
conda activate bioinfo
```

The bulk RNA-Seq test data we've been working with is in FASTQ format. We'd like to do a BLAST search on a couple of these sequences. Data must be in FASTA format to do BLAST. How can we transform FASTQ to FASTA format?

We will continue working in the directory we already created for this data.

```
cd biostar_class
cd RNA_Seq
ls
```

We can use the EMBOSS Seqret program. This tool is also available on the web where you can copy and paste sequences into a graphical user interface (GUI). For working with very large files however, we will want to work at the command line. To use the "seqret" tool, the command line must be in this format, where "-sequence" is followed by the name of the input sequence, and "-outseq" is followed by the name of the output file.

```
seqret -sequence reads.fastq -outseq reads.fasta
```

To see all the options available for the "seqret" program, just type this at the command line.

```
seqret --help
```

and to see even more...

```
seqret --help --verbose
```

In this example, "reads.fastq" is the input file and "reads.fasta" is the output file. Let's try this on some of our bulk RNA-Seq test data. First, we need to "gunzip" one of the files, as "seqret" can not work with compressed files. Let's go ahead and gunzip one of the smaller ".fastq.gz" files. Check the file sizes of the fastq files with

```
ls -l
```

Now do the "gunzip" step on one of the smaller files.

```
gunzip HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-
chr22.read1.fastq.gz
```

Note the size of this file before (7.3 M) and after (30M) doing the "gunzip" command. We can look inside this file with the "less" command. By "piping" it through the head command again, we can see just the first eight lines of the file.

```
less HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-
chr22.read1.fastq | head -n 8
```

to see the results

```
@HWI-ST718_146963544:8:1212:5958:93757/1
CTGATTTAACAGCAACACATTTGACTTCAATATTATGGCAAATGAGTCCAAAGTCATACTGCCC

+
CCCFFFFFHHHHHJJIIJJJJJJJJJJJJJJIIJIJJJJIJJJJJJIIJJJJJHHIJJJIIIJJ

@HWI-ST718_146963544:7:2308:7250:88065/1
CGTCGATGTATGCACTCATTATTAGATCCTCAGTATGTATGGTTTCAGCTATGAATGAAAGCAT

+
@C@DFFFAFHGFHIIIIAHIIIIIIIIGIIEHGHHGG<GGACHDHGCHGGEIIIIIIIIIIFII
```

Now we can run the "seqret" command to transform the .fastq file to .fasta

```
seqret -sequence HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-
chr22.read1.fastq -outseq HBR_Rep3_ERCC-Mix2_Build37-
ErccTranscripts-chr22.read1.fasta
```

Let's peek inside our .fasta output file with cat, like we did with the compressed .fastq file. This time we can use the "cat" command, since the file is not compressed. (The "less" command would also work here.)

```
cat HBR_Rep3_ERCC-Mix2_Build37-ErccTranscripts-
chr22.read1.fasta | head -n 8
```

Now we can see the first eight lines of the .fasta file.

```
>HWI-ST718_146963544:8:1212:5958:93757/1
CTGATTTAACAGCAACACATTTGACTTCAATATTATGGCAAATGAGTCCAAAGTCATACT
GCCCCTCCCGTTTGGTATTTTAAGTACAGTAATAGACAAA
>HWI-ST718_146963544:7:2308:7250:88065/1
CGTCGATGTATGCACTCATTATTAGATCCTCAGTATGTATGGTTTCAGCTATGAATGAAA
GCATTGCCTCCTTTCTTGTTGACCTGAGTTTACTAAGTAA
>HWI-ST718_146963544:8:1210:13422:9826/1
AGTTGGGGTCCTAAGCCAGAAGTTAACTATGCTTCATATATTCTTGCAAGTAGAAGTACA
```

Another useful tool for working with FASTQ files is seqkit. There are many examples of how to use this tool in the Biostar handbook, we'll review a couple of them here. To see all the available commands for "seqkit"...

```
seqkit --help
```

Here is a selection of "seqkit" commands.

```
Usage:
  seqkit [command]

Available Commands:
```

```
  amplicon        retrieve amplicon (or specific region
around it) via primer(s)
  bam             monitoring and online histograms of BAM
record features
  common          find common sequences of multiple files by
id/name/sequence
  concat          concatenate sequences with same ID from
multiple files
  convert         convert FASTQ quality encoding between
Sanger, Solexa and Illumina
  duplicate       duplicate sequences N times
  faidx           create FASTA index file and extract
subsequence
  fish            look for short sequences in larger
sequences using local alignment
  fq2fa           convert FASTQ to FASTA
  fx2tab          convert FASTA/Q to tabular format (with
length/GC content/GC skew)
  genautocomplete generate shell autocompletion script
  grep            search sequences by
ID/name/sequence/sequence motifs, mismatch allowed
  head            print first N FASTA/Q records
  help            Help about any command
  locate          locate subsequences/motifs, mismatch
allowed
  mutate          edit sequence (point mutation, insertion,
deletion)
  range           print FASTA/Q records in a range
(start:end)
  rename          rename duplicated IDs
```

Let's retrieve the data as shown in the Biostar Handbook (Advanced FASTQ processing). We'll use the "curl" command instead of "wget", and specify an output file using "-o" (output).

```
curl http://data.biostarhandbook.com/reads/duplicated-
reads.fq.gz -o duplicated-reads.fq.gz
```

Do this for the next two data sets. Here's something a little different, we're retrieving from an "ftp" site (ftp = file transfer protocol) instead of an "http" site.

```
curl
ftp://ftp.ncbi.nih.gov/refseq/release/viral/viral.2.1.genomic.fn
 -o genomic.fna.gz
```

and

```
curl
ftp://ftp.ncbi.nih.gov/refseq/release/viral/viral.2.protein.faa.
 -o protein.faa.gz
```

Let's run the "stat" function to get some information about these files we've downloaded.

```
seqkit stat *.gz
```

Here are the results.

```
file                    format  type     num_seqs
sum_len  min_len   avg_len    max_len
duplicated-reads.fq.gz  FASTQ   DNA        15,000
```

```
1,515,000      101        101         101
genomic.fna.gz           FASTA   DNA          8,431
212,623,662      200  25,219.3  2,243,109
protein.faa.gz           FASTA   Protein   247,311
64,386,104        7     260.3      8,573
```

Here's how we can get the GC content of those files.

```
seqkit fx2tab --name --only-id --gc *.gz | head
```

And here are the results.

```
SRR1972739.1                    30.69
SRR1972739.2                    49.50
SRR1972739.3                    41.58
SRR1972739.4                    51.49
SRR1972739.5                    48.51
SRR1972739.6                    31.68
SRR1972739.7                    38.61
SRR1972739.8                    36.63
SRR1972739.9                    40.59
SRR1972739.10                   31.68
```

+ Add a custom sidebar

Clone this wiki locally

https://github.com/AmyStonelake/BTEP.wiki.git