Quality Control of Sequencing Data Solutions

Surya Saha

Sol Genomics Network (SGN)

Boyce Thompson Institute, Ithaca, NY

ss2489@cornell.edu // Twitter:@SahaSurya

BTI Plant Bioinformatics Course 2017

Exploration

Goal:
Use shell commands and installed tools to explore using the file you downloaded from TAIR

Data:
/home/bioinfo/Data/TAIR10_pep.fasta.gz

Tools:

EXONERATE (fasta toolkit)

Exploration

Exercise 1:

1. Type fasta and TAB key to find all the commands starting with fasta

2. Uncompress and find the length of sequences in the file: TAIR10_pep.fasta.gz

```
gunzip TAIR10_pep.fasta.gz
```

```
fastalength -f TAIR10_pep.fasta
```

3. Split the above file into 3 fasta files

```
fastasplit -f TAIR10_pep.fasta-c 3 -o /home/bioinfo/Data
```

Evaluation

Goal:

Learn the use of read evaluation programs keeping attention in relevant parameters such as qscore and length distributions and reads duplications.

Data:

(Illumina data for two tomato ripening stages)
/home/bioinfo/Data/Slch04_demo.tar.gz

Tools:

tar -zxvf OR –xvf (command line, untar and unzip the files)
head (command line, take a quick look of the files)
grep (command line, find/count patterns in files)
FASTX toolkit (command line, process fasta/fastq)
FastQC (gui, to calculate several stats for each file)

Evaluation

Exercise 2:

1. Uncompress the file:
/home/bioinfo/Data/test_RNAseq.tar.gz

2.Print the first 10 lines for the files: SRR404331_ch4.fq, SRR404333_ch4.fq, SRR404334_ch4.fq and SRR404336_ch4.fq.

Question 2.1: Do these files have fastq format?

Evaluation

Solution 2:

1. Untar the file: *ch4_demo_dataset.tar.gz*

```
tar -xvf Slch04_demo.tar.gz
```

2. Raw data will be found in 4 files. Print the first 10 lines for the files: SRR404331_ch4.fq, SRR404333_ch4.fq, SRR404334_ch4.fq and SRR404336_ch4.fq.

```
head SRR404331_ch4.fq; head SRR404333_ch4.fq; head SRR404334_ch4.fq; head SRR404336_ch4.fq;
```

Question 2.1: Do these files have fastq format? Yes

Evaluation

Solution 2:

3. Count number of sequences in each fastq file using commands you learnt last time

```
grep -c '^+$' SRR404331_ch4.fastq
```

Evaluation

Solution 2:

4. Convert the fastq files to fasta

```
fastq_to_fasta -I SRR404331_ch4.fastq -o SRR404331_ch4.fasta -Q33 -n
```

The –Q33 flag is to denote Sanger Phred 33 encoding. It expects Illumina Phred+64 by default. See http://seqanswers.com/forums/showthread.php?t=7596

-n tells it not to remove any sequences from the file. It removes any reads containing N's by default

5. Now count the number of sequences in fasta file and see if the number of sequences has changed.

```
grep -c '^>' SRR404331_ch4.fasta
```

It may be different if you did not use -n parameter. Not using -n will remove sequences with any 'N' bases

Evaluation

Solutions 3:

Question 3.2: How many sequences there are per file in FastQC?

SRR404331_ch4.fastq = 762,365

SRR404333_ch4.fastq = 744,048

SRR404334_ch4.fastq = 592,123

SRR404336_ch4.fastq = 880,982

Question 3.3: Which is the length range for these reads?

53 in most of the cases, except in SRR404336 where is 54

Question 2.4: Which is the qscore range for these reads? Which one looks best quality-wise?

The range goes from 2 in some cases such as SRR404334 to a maximum of 44. SRR404331 has consistent quality over entire length with least 3' end drop-off.

Evaluation

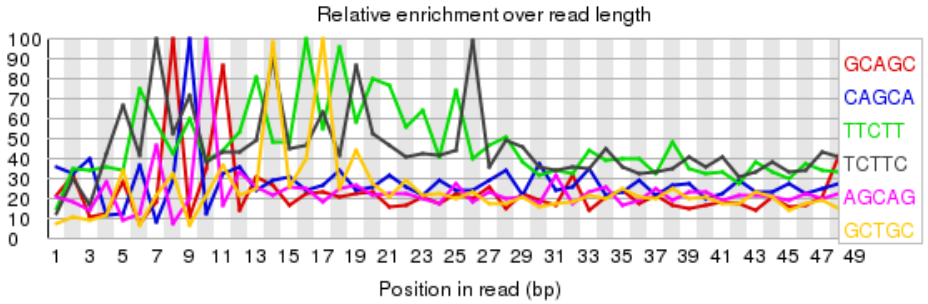
Solutions 3:

Question 3.5: Do these datasets have read overrepresentation?

In some cases such as SRR404331. A Blast search of the top overrepresented sequence reveals that it is the gene ASR1 from tomato, so it is not a contamination and probably it has some biological relevance.

Question 3.6: Looking into the kmer content, do you think that the samples have an adaptor?

No, it is possible to see some structure at the 5' extreme, but it starts at the position 7 and doesn't affect to all the sequences.



Preprocessing

Goal:
Trim the low quality ends of the reads and remove the short reads.

Data:
(Illumina data for two tomato ripening stages)
ch4_demo_dataset.tar.gz

Tools:
fastq-mcf (command line tool to process reads)
FastQC (gui, to calculate several stats for each file)

Preprocessing

Exercise 4:

- Download the file: adapters1.fa from ftp://ftp.solgenomics.net/user_requests/aubombarely/courses/RNAseqCorpoica/adapters1.fa
- Run the read processing program over each of the datasets using a min. quality score of 30 and a min. length of 40 bp.

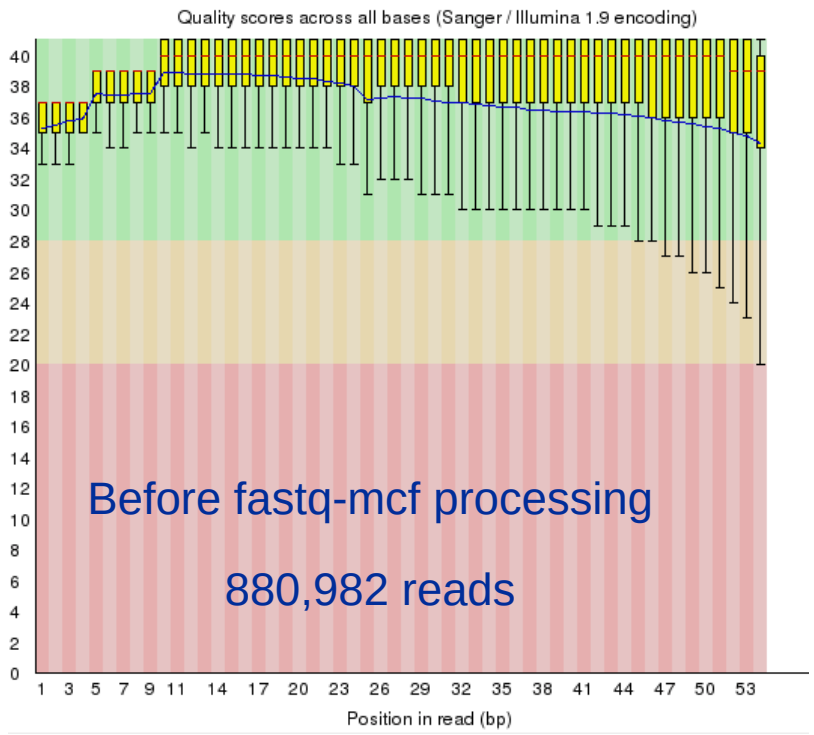
```
fastq-mcf -q 30 -l 40 -o SRR404331_ch4.q30l40.fastq
/home/bioinfo/Downloads/adapters1.fa SRR404331_ch4.fastq
```

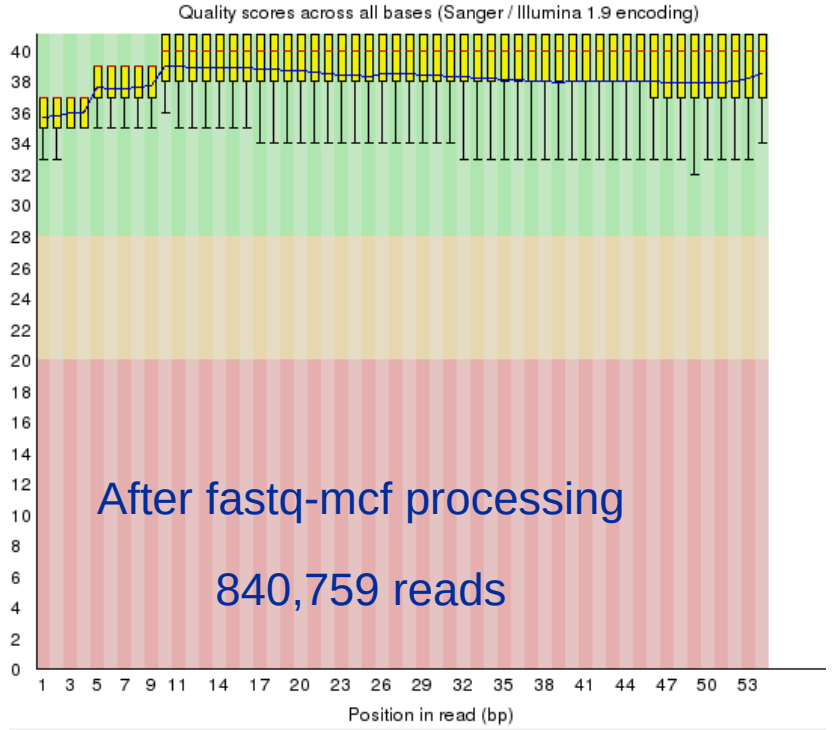
Remember to specify the right path for the adapter1.fa and the input file

Preprocessing

Exercise 4:

•Type ‘fastqc’ to start the FastQC program. Load the four sequence files in the program. Compare the results with the previous datasets.





SRR404336

Bonus Material

Here are some simple exercises using the file from TAIR /home/bioinfo/Data/TAIR10_pep.fasta.gz

- Number of unknown proteins per chromosome
- Number of proteins that are not unknown proteins per chromosome and are on the forward strand
- Number of proteins on chromosome 1 which are located within the first 5MB of the chromosome
- All proteins of length greater than 200aa in chromosome 1

Bonus Material

- Number of unknown proteins per chromosome

```
grep '^>' TAIR10_pep.fasta | grep  'unknown protein' |cut -c 1-4| sort | uniq –c
   934 >AT1
   632 >AT2
   764 >AT3
   550 >AT4
   831 >AT5
```

- Number of proteins that are not unknown proteins per chromosome and are on the forward strand

```
grep '^>' TAIR10_pep.fasta | grep -v 'unknown protein' |grep 'FORWARD ' | cut -c 1-4| sort
| uniq –c
  3053 >AT1
  1821 >AT2
  2273 >AT3
  1813 >AT4
  2727 >AT5
    32 >ATC
    45 >ATM
```

Bonus Material

- Number of proteins which are located within the first 5MB of the chromosome

```
grep '^>' TAIR10_pep.fasta | cut -d '|' -f 4| cut -d '-' -f 2| cut -d ' ' -f 1| awk '{if ($1 < 5000000) print}'| wc -l
```

5864

- All genes of length greater than 2000bp

```
grep '^>' TAIR10_pep.fasta | cut -d '|' -f 4| awk '{print $1}'| awk -F":" '{print $2}'| awk -F"-" '{if($2-$1 > 2000) print}'| wc -l
```

9647

- All proteins of length greater than 200aa

```
grep '^>' TAIR10_pep.fasta | cut -d '|' -f 4| awk '{print $3}'| awk -F"=" ' $2 > 200 {print }' | wc -l
```

20413