Tutorial 3

Next generation sequencing
Mohamed Hamed

August 27, 2014

Note: As S. aureus reference genome, NC_017340 (04-02981) should be used.
http://www.ncbi.nlm.nih.gov/nuccore/NC_017340.1

Getting ready
-Copy the NGS data from *\/home/stud/mhamed /Tut03* to your home directory

Exercise 1. Quality Control of fasta files

-Open a terminal on your computer and type the command fastqc
-Load the four sequence files existing in *Tut03* into fastqc program

-Which platform were they produced with?
-How many reads are in each sequence file?
-What is the read length?
-Comment on the quality of each sequence file as mentioned in the lecture. Per base sequence quality, per sequence quality scores,..etc
Hints:

1. http://en.wikipedia.org/wiki/FASTQ_format
2. For saving disk space for the next steps you might delete the two sequence files: S12_R1.fastq, and S12_R2.fastq

Exercise 2. Allignment using BWA

2.1 creatng a bwa index from the reference file

- Download the reference genome NC_017340 from NCBI in a fasta format and save as "ref.fasta" into the same directory of your sequencing data.

-Open a terminal and go to your data directory

```
cd Tut03
```

Then type the following command to create an index for the reference genome

```
bwa index -a is -p MRSA17340   ref.fasta
```

An index for the reference with name MRSA17340 will be created.

2.2 map the two pairs and join in one sam file

```
bwa aln  MRSA17340 R4_R1.fastq > R4_R1.sai
bwa aln  MRSA17340 R4_R2.fastq > R4_R2.sai
bwa sampe MRSA17340 R4_R1.sai R4_R2.sai R4_R1.fastq  R4_R2.fastq > R4.sam
```

2.3 convert to a bam file and sort and index

```
samtools view -bSh  R4.sam | samtools sort - R4s
samtools faidx ref.fasta
samtools index R4s.bam
```

2.4 sam file statistics

```
samstat R4.sam
```

OR

```
samstat R4s.bam
```

-How many reads are mapped with mapping quality over 30?
-What are the most over-represented 10-mers in the mapped reads with quality over 30?

use samtools manual and descriptions (http://samtools.sourceforge.net/samtools.shtml) to answer the following:

-How many reads are
…….in the bam file?

```
samtools view R4s.bam | wc -l
```

……..mappedin total?

```
samtools view -F 4 R4s.bam | wc -l
```

……..mapped with min mapping quality equal to 30?

```
samtools view -q30 -b R4s.bam | wc -l
```

……..paired in sequencing?

```
samtools view -f 1 R4s.bam | wc -l
```

…….read 1?

```
samtools view -f 64 R4s.bam | wc -l
```

……..properly paired?

```
samtools view -f 2 R4s.bam | wc -l
```

……..duplicates?

```
samtools view -f 1024 R4s.bam | wc -l
```

2.5 filter the reads having mapping quality less than 30 and remove the duplicate reads existed by pcr

```
samtools view -q30 -b R4s.bam | samtools rmdup - R4sfr.bam
```

Check again the no of reads in the file R4sfr.bam and compare it with the no of reads you obtained above (reads with quality over 30) the same

```
Samtools view  -b R4sfr.bam | wc -l
```

2.6 index the final bam file

```
samtools faidx ref.fasta
samtools index R4sfr.bam
```

Exercise 3. Looking at alignment files in graphical view

Examine the BAM file with samtools tview and IGV.

3.1-samtools tview

```
samtools tview R4sfr.bam
```

-Are the reads sorted?. Press "?" To access the help and get an overview for the available commands

3.2-Integrative genome browser (IGV)

-Open IGV by writing the following command

```
igv &
```

-Load the genome located in the data folder and named "MRSA017340.genome"
-Make sure that the reference genome is set to "MRSA017340".
-Load the bam file "R4sfr.bam"
-Examine the position of the SNP on position 83067
-What is the genotype?
-How many reads are aligned there?
What are their mapping qualities?
What are the base qualities of the variant bases?
What is the name of the gene (or the locus tag) where SNP occur?

Similarly check the positions: 122203, 246160, and 907863

Exercise 4. Assembly

The exercise will show you how to perform de-novo assembling in case you don't have (or can't map reads to) reference genome.
In this tutorial, you will learn how to use SPAdes assembler to construct contigs from unmapped reads of *S. aureus.*

-Go to the assembly irectory where you have the unmapped reads

```
cd assembly
```

And then provide the input file that contains interlaced reads (both forward and reverse reads) as denoted in the following command

```
spades --12 [interlaced_read_file] -o [output_dir]
```

Where the interlaced reads file is: S11_unmapped.fastq and the output directory is " results". When the job is finished, all the output files should be stored in `[output_dir]`. The `contigs.fasta` in `[output_dir]` contains resulting contigs.
- Generate a ref genome from the contigs file by using the IGV file-import genome and add refer to the "contigs.fasta" file