Exercises NGS tools Tutorial

Download the files:
https://www.dropbox.com/s/oe5okkerl085k4y/NGStools.zip

Work locally on your laptop
The following instructions have been tested on Ubuntu. Mac and Windows users can use Ubuntu in the Virtual Machine (setup distributed by Stefan Wyder)

1. Copy files to your computer:
```
wget https://www.dropbox.com/s/oe5okkerl085k4y/NGStools.zip
```

2. Install seqtk:
```
wget https://github.com/lh3/seqtk/archive/master.zip
unzip master.zip
cd seqtk-master
make
```

3. Install vcftools:
```
sudo-apt get install vcftools
```

   or

```
wget https://github.com/vcftools/vcftools/zipball/master
unzip master
cd vcftools-vcftools-78add55
./autogen.sh
./configure
make
sudo make install
```

4. Install bcftools:
```
wget
https://github.com/samtools/bcftools/releases/download/1.2/bcftoo
ls-1.2.tar.bz2
tar –xjvf bcftools-1.2.tar.bz2
cd bcftools-1.2
make
sudo make install
```

Exercises FASTQ files:

The downloaded zip file contains two FASTQ files from a human sample of the 1000 genomes project. It is a paired-end data set with an insertion size of 200bp.

1. Make a quality check of the FASTQ files

2. Imagine you tried to map the reads to the human reference genome and following reads failed the mapping process:

```
ERR000064.19 BGI-FC20CNLAAXX_7_1_944:632/1
ERR000064.336257 BGI-FC20CNLAAXX_7_14_339:916/1
ERR000064.346608 BGI-FC20CNLAAXX_7_14_977:474/1
ERR000064.521728 BGI-FC20CNLAAXX_7_22_625:861/1
ERR000064.521729 BGI-FC20CNLAAXX_7_22_380:228/1
ERR000064.602263 BGI-FC20CNLAAXX_7_25_860:760/1
ERR000064.406739 BGI-FC20CNLAAXX_7_17_574:189/2
ERR000064.406740 BGI-FC20CNLAAXX_7_17_533:077/2
ERR000064.521729 BGI-FC20CNLAAXX_7_22_380:228/2
ERR000064.534039 BGI-FC20CNLAAXX_7_22_400:333/2
```

   Extract the above sequences from the FASTQ files and save them in one FASTA file. Theoretically, this file could then be used in a BLAST search to check for contamination.

3. Some programs require that paired-end reads is stored in one interleaved FASTQ files. Thus convert the two FASTQ files into one interleaved FASTQ file.

4. Randomly down sample the FASTQ files to 1000 paired-end sequences.

Exercises VCF file:

The downloaded zip file contains a VCF file with variant positions on chromosome 18 and 19 between 60 human samples (30 British and 30 South Han Chinese samples) of the 1000 genomes project (phase 3).

1. Create a new VCF file with all SNPs located on chromosome 19 between position 1 and 10,000,000.

2. Out of the new VCF file create two VCF files one for each population and only keep SNPs polymorphic within these populations
   (the sample files for each population can be found in the zip file: British.txt and HanChineseSouth.txt).

3. Compare the SNPs between the two populations

4. Merge the two population VCF files into one

5. Create a consensus sequence of chromosome 18 for each population and reverse complement it.

6. Estimate Fst between the two populations and nucleotide diversity within each population in a sliding window along chromosome 18 (window size: 100 kb, step size: 50 kb).
   Use R to plot the Fst and nucleotide diversity along chromosome 19.