Week 4: Nextgen Sequence Quality Control

Molb 4485/5485 -- Computers in Biology

1. Login to the Teton Environment

- Log onto Teton using your account, password, and Yubikey token as we did for the last 3 weeks:

```
ssh <YourNetID>@teton.uwyo.edu
```

- remember ALL PASSWORDS are **wyoinbre,YUBIKEY-TOKEN**
- Let's all move over to the working directory we have used for the past 3 weeks:

```
cd /project/inbre-train/<yourNetID>/

pwd
```

- You should see `/project/inbre-train/<YourNetID>`, *e.g.*, `inbre001`.
- Just like last week, the materials you will need are in the directory for this week (week 4):

```
ls -l /project/inbre-train/

drwxrwsr-x 6 inbre001 inbre-train 4096 Oct 23 14:05 inbre001
drwxrwsr-x 5 inbre002 inbre-train 4096 Oct 23 14:04 inbre002
drwxrwsr-x 6 inbre003 inbre-train 4096 Oct 23 14:03 inbre003
...
...
...
drwxr-sr-x 3 nblouin  inbre-train 4096 Oct 23 14:03 nblouin
drwxr-sr-x 7 vchhatre inbre-train 4096 Oct 26 10:30 vchhatre
drwxrwsrwx 2 nblouin  inbre-train 4096 Oct 15 14:23 Week2Data
drwxrwsrwx 2 nblouin  inbre-train 4096 Oct 23 16:19 Week3Data
drwxrwsrwx 3 nblouin  inbre-train 4096 Oct 29 13:21 Week4
```

- Now we need a directory where we will do all of our work for this week.

```
mkdir <your_last_name>_Week4    ## Example: Blouin_Week4
```

- Move into that directory you just made (`cd`):

```
cd Blouin_Week3
```

- Copy (**cp**) the paired-end DNA read files into your working directory from `/project/inbre-train/Week4`. These data are from the 1000 Genomes Project. Part of the data you will prepare this week will be used to

investigate genetic variation between different human populations in the context of their geographic position (last class meeting).

2. Quality Control

The first thing many people want to know is: Are my sequences garbage?

Lots can happen after you send in your precious sample for sequencing. Accidents happen all the time.

While some informatics programs can take raw reads (untrimmed data) as input, it is good practice to do this yourself before running any analysis.

These will serve as a basic guideline for read prep. This assumes you have demultiplexed Illumina data. If you ever run across a multiplexed dataset, write us and we will help you with that more advanced task.

2.1 FastQC Sequence Quality Assessment

Today you will see FastQC runs easily on the command line. There is also a GUI version of this program, but it has its limits. When you have LARGE datasets that are difficult to move, you will quickly come to love FastQC on the command line so you can assess your datasets on a computer cluster (*e.g.*, Teton).

When doing quality control (QC) you must evaluate each paired read file separately. Do not assume that each read file is OK if you only checked one of two files. In most cases the reverse reads (read 2 files) will be of poorer quality than the matching forward reads.

1. Now that you have copied (cp) the paired reads files for `ERR` in the Week4Data directory to your current working directory here are the files:

```
ls -othr /project/inbre-train/Week4/
```

```
-rwxrwxrwx 1 nblouin 1.9G Oct 29 12:48 ERR013161_2.filt.fastq.gz
-rwxrwxrwx 1 nblouin 1.9G Oct 29 12:48 ERR013161_1.filt.fastq.gz
```

These file are compressed (`*gz`). Some programs require you to unpack them. Lucky for us that the modules (programs) we are using today will do that in the background for us.

2. Run FastQC. You can get help by running the following command

```
module load gcc swset fastqc
```

3. Take a quick peek at the usage for FastQC (`-h`)

```
fastqc -h
```

Like last week, we will send a job to the Teton computing environment. To do this for QC assessment of our sequences we will submit the following script. Make a shell a script called `FastQC_Before.sh using the text editor (vi). What does the`–t 4“ command do in the shell below?

```
#!/bin/bash
#SBATCH -J FQC_Pre
#SBATCH -n 1
#SBATCH -t 30:00
#SBATCH --cpus-per-task=4
#SBATCH --mail-type=ALL
#SBATCH --mail-user=<your email address>
#SBATCH --account=inbre-train

echo "Loading Modules"

module load  swset gcc
module load fastqc

echo "Following modules have been loaded"
module list

echo "Initiating FastQC Run at $(date)"

fastqc -t 4 ERR013161_1.filt.fastq.gz ERR013161_2.filt.fastq.gz


echo "Finished FastQC Run at $(date)"
```

4. Each read set will create two output files `<read_name>.fastqc.html` and `<read_name>.fastqc.zip`. Take a look

```
ls -othr
```

5. Now you need to download the html files and look at them using a web browser. Remember in Week2 and we used the secure copy command (scp) to download files to you workstations? Remember we do not type "<" and ">". The syntax for secure copy is:

```
scp <user@server>:<PathToWhatIWantToCopy> <WhereIWantToCopy>
```

To do this open a new terminal window (this is a local window NOT connected to Teton) and copy the html files like this all in one line, start by getting your working directory for the path to the files:

```
pwd

# This gives us the path we will need for the command below, mine looks like this
# /project/inbre-train/nblouin/Blouin_Week4
```

```
scp <YourUserName>@teton.uwyo.edu:/project/inbre-train/<YourUserName>/<YourLastName>_Week4/*
```

6. Open The HTML files on your desktop and let's look at the datasets.

Based upon this output you can make decisions about the read quality and what needs to be trimmed. Typically each dataset will have it's own parameters, but each pair will share parameters.

2.2 Sequence Quality Control with Trimmomatic

There are a lot of trimming tools. The one we will use today is called Trimmomatic. You can see its user manual by clicking on this link.

We need to decide how we want to trim. Note Trimmomatic trims in the order the commands are given.

In the case of the ERR reads from the 1000 Genomes (Human) project, we will want to trim the first 15 bps from the 5'-end (beginning of the sequence) as well as low quality reads. Then we will remove any remaining short sequences.

Note: Trimmomatic will remove duplicates as well. Normally, for genome assembly this is something that we would do. However in the interest of time we will skip this step. Please review the Trimmomatic manual for all of the settings and suggestions.

Now we can build our job command. Below you will see a new symbol (\) that allows us to use line breaks. Whenever the computer sees this symbol \, it knows you are finishing your command on the following line. After each \ hit the enter key and keep typing. Let's make a batch script (also called a shell or shell script) using your text editor (`vim`). Call this shell script `Trim.sh`

```
#!/bin/bash
#SBATCH -J Trim
#SBATCH -n 1
#SBATCH --cpus-per-task=4
#SBATCH -t 30:00
#SBATCH --mail-type=ALL
#SBATCH --mail-user=<your email address>
#SBATCH --account=inbre-train

echo "Loading required modules"

module load swset gcc trimmomatic

echo "Following modules have been loaded:"
module list
```

```
echo "Initiating Trimmomatic Run at $(date)"

trimmomatic PE \
    -threads 4 \
    ERR013161_1.filt.fastq.gz ERR013161_2.filt.fastq.gz \
    ./fwd_pair.fq ./fwd_unpair.fq ./rev_pair.fq ./rev_unpair.fq \
    HEADCROP:10 \
    SLIDINGWINDOW:4:24 \
    MINLEN:80

echo "Completed Trimmomatic Run at $(date)"
```

2.3 Task: Compare Trimmed Reads to Raw, Un-Trimmed Reads

1. Run FastQC on the trimmed paired-end reads (only the two "pair" files i.e. `fwd_pair` and `rev_pair`). You may ignore the `unpair` files.
2. Download the html files (`scp`) and look at them. What are at least 3 things that have changed? **We encourage you to chat with your neighbors.**
3. Look in your `slurm.out` file to see how many reads were removed in your trim. What are some different choices, if any, you might make if this were your data? You might need to look at the **User Manual for Trimmomatic** for ideas.