Writing Assignment 2 - BINF 8940

K. Bodie Weedop

2021/02/10

Notes:

Accessing teaching cluster: `ssh MyID@teach.gacrc.uga.edu`

All data necessary for this list is available at: `/work/binf8940/instructor_data/WA2_files`

Illumina PE reads:

- `/work/binf8940/instructor_data/ WA2_files/Ecoli_short_1.fastq`
- `/work/binf8940/instructor_data/ WA2_files/Ecoli_short_2.fastq`

ONT Reads:

- `/work/binf8940/instructor_data/WA2_files/Ecoli_ONT.fastq`
- `/work/binf8940/instructor_data/WA2_files/Ecoli_ONT.fasta`

Reference genome: `/work/binf8940/instructor_data/WA2_files/Ecoli_reference.fasta`

Please use the GACRC software wiki to properly run jobs in the teaching cluster. ([https://wiki.gacrc.uga.edu/wiki/Software])

Before starting any genome assembly, it is always important to trim the raw reads to avoid low quality reads and remove adapters. Trimmomatic is a well known tool to do so.

Example command:

```
ml Trimmomatic/0.39-Java-11
java -jar $EBROOTTRIMMOMATIC/trimmomatic-0.39.jar PE -phred33 File_1.fastq
File_2.fastq output_1_paired.fq output_1_unpaired.fq output_2_paired.fq
output_2_unpaired.fq ILLUMINACLIP:/apps/eb/Trimmomatic/0.39-Java-
11/adapters/TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20
MINLEN:40
```

- Run trimmomatic using your Illumina reads in the cluster
- How many of these reads were truly paired? (P1+P2)
  - 2213032
- If you change SLIDINGWINDOW:4:20 to SLIDINGWINDOW:4:30, what will happen? Was there any change in your result?
  - By changing the value from 20 to 30, you are increasing the average quality required and, therefore, increasing the threshold for the cuts.

    - There was a change in the result. There are now only 1603653 reads that were paired
- Paste one of your teaching cluster shell scripts below.

```
#!/bin/bash
#SBATCH --job-name=weedopTrimmomatic
#SBATCH --partition=batch
#SBATCH --mail-type=ALL
#SBATCH --mail-user=kbw81711@uga.edu
#SBATCH --ntasks=1
#SBATCH --mem=10gb
#SBATCH --time=08:00:00
#SBATCH --output=Trimmomatic.%j.out
#SBATCH --error=Trimmomatic.%j.err

# Illumina reads
illumina1=/work/binf8940/instructor_data/WA2_files/Ecoli_short_1.fastq
illumina2=/work/binf8940/instructor_data/WA2_files/Ecoli_short_2.fastq

# Load module
ml Trimmomatic/0.39-Java-11
# Run command
java -jar $EBROOTTRIMMOMATIC/trimmomatic-0.39.jar PE -phred33 $illumina1
$illumina2 output_1_paired.fq  output_1_unpaired.fq output_2_paired.fq
output_2_unpaired.fq ILLUMINACLIP:/apps/eb/Trimmomatic/0.39-Java-
11/adapters/TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:30
MINLEN:4
```

Now that you have your trimmed reads, lets learn how to make a reference genome assembly. The commands below are an example of how to generate an alignment of your reads against the reference:__

Bwa mem:

Example command:

```
ml SAMtools/1.9-GCC-8.3.0 BWA/0.7.17-GCC-8.3.0
bwa index Reference_Genome.fasta
bwa mem Reference_Genome.fasta File_1.fastq File_2.fastq >
out_algn_mem.sam
samtools view –b –S out_algn_mem.sam > out_algn_mem.bam
samtools sort -o out_algn_mem.sorted.bam out_algn_mem.bam
samtools index out_algn_mem.sorted.bam
```

To generate the Reference based assembly here is an example:

```
ml BCFtools/1.10.2-GCC-8.3.0
bcftools mpileup -f ref.fa sample.sorted.bam | bcftools call -mv -Oz -o calls.vcf.gz
tabix calls.vcf.gz
cat ref.fasta | bcftools consensus calls.vcf.gz > consensus.fasta
```

Run the Reference based using the example above with the data used in WA1.

- How many sequences were generated in your Consensus.fasta file?
  - 1
- What is the size of the genome assembled (use UNIX commands)?
  - 4718920 bp

```
cat consensus.fasta | grep -v ">" | wc -m
```

De novo assembly (Illumina short reads):

Run SPADES for a de novo genome assembly:

- Which was the best K-mer (you can try using the -k auto from spades or Kmergenie in the cluster)
- How many contigs/scaffolds did you find after spades run?
- Paste your shell script below.

De novo assembly (ONT):

- Perform the ONT de novo assembly. Run Flye (ml Flye/2.6-foss-2019b-Python-3.7.4) with default parameters for --nano-raw (paste your script – Probably your run will not be finished today)

  - Usage: flye --nano-raw --genome-size (Check on NCBI for E.coli) --out-dir Ecoli_Flye

- Polish using Racon (works with both Illumina and Long-reads) and Pilon (Illumina):

- Polish using Just ONT reads

  - Example:

    ```
    ml minimap2/2.17-GCC-8.3.0
    # Align reads to the assembled genome
    minimap2 -x map-ont --secondary=no -t 10 ONT_consensus_miniasm.fasta ONT_reads.fasta > polish_miniasm.paf
    # Polish the genome by finding consensus of aligned reads at each position
    ml Racon/1.4.13-GCCcore-8.3.0
    racon -t 10 ONT_reads.fasta polish_miniasm.paf ONT_consensus_miniasm.fasta > Ecoli_nano_Racon.fasta
    ```

- Polish using Illumina reads: (remember to align your short reads against the Flye assembly result using BWA)
    - Example:

      ```
      ml Pilon/1.23-Java-11
      java -Xmx70G -jar --genome Flye_scaffods.fasta --bam
      Ecoli_short_reads.sorted.bam --output prefix --changes   --
      threads 10
      ```

What are the differences that you believe will be observed of these generated assemblies? Why and how it could impact a future analysis?

Keep all results for the Validation Class!