Starling-May18
Projects/Katarina

PDF Version generated by

Katarina Stuart (z5188231@ad.unsw.edu.au)

on

Jun 23, 2022 @03:53 PM NZST

Table of Contents

2020-03-05.Pilon.ScaffoldedGenome

Katarina Stuart (z5188231@ad.unsw.edu.au) - Jan 11, 2021, 3:58 PM NZDT

Pilon on the scaffolded genome

http://protocols.faircloth-lab.org/en/latest/protocols-computer/assembly/polishing-with-pilon.html

```
RAWDATA=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/rawdata/HN00105164/HN00105164_10x_RawData_Outs/H2CYFCCX2/fastq_path/H2CYFCCX2/SV01
```

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/Pilon/supernova_raw
```

```
ln -s $RAWDATA/SV01_S1_L006_R1_001.fastq.gz SV01_S1_L006_R1_001.fastq.gz
ln -s $RAWDATA/SV01_S1_L006_R2_001.fastq.gz SV01_S1_L006_R2_001.fastq.gz
```

```
mkdir longranger-ouput && cd $_
```

1) Use longranger to process the reads.
This should take <24 hours for ~40 GB zipped sequence data. The processing basically trims the reads to remove the barcode and adapter information and puts the barcode info in the fastq header:

```
#!/bin/bash

#PBS -N 2020-02-04.Pilon
#PBS -l nodes=1:ppn=24
#PBS -l vmem=24gb
#PBS -l walltime=48:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/Pilon/longranger-ouput

module add longranger/2.2.2

longranger basic --id=SV01 --fastqs=../supernova_raw/ --localcores 24 1>longranger-basic.stdout 2>longranger-basic.stderr
```

This has been done for the first version of the genome already (non-scaffolded)

2) Map reads to genome assembly
Once the reads have been processed, we want to map them to our genome assembly using bwa-mem and samtools.

```
mkdir bwa-aligned_scaffolded && cd $_

#symbolic link of the longeranger output from above
ln -s ../longranger-ouput/SV01/outs/barcoded.fastq.gz
```

```
ln -s /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/L_RNA_scaffolder/L_RNA_scaffolder.fasta
```

```
#!/bin/bash
```

```
#PBS -N 2020-03-05.PilonMap_scaffolded.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=350gb
#PBS -l walltime=24:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae
```

```
module load bwa/0.7.17
module load samtools/1.10
module load java/8u45
module load pilon/1.23
```

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/Pilon/bwa-aligned_scaffolded
```

```
# index the assembly for bwa
bwa index L_RNA_scaffolder.fasta
```

```
#!/bin/bash

#PBS -N 2020-03-12.PilonAlign_scaffolded.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=56gb
#PBS -l walltime=48:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

module load bwa/0.7.17
module load samtools/1.10
module load java/8u45
module load pilon/1.23

cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/Pilon/bwa-aligned_scaffolded

# run bwa, use 20 threads for aligning and sorting and set memory for samtools at 3G per thread
bwa mem -t 14 L_RNA_scaffolder.fasta barcoded.fastq.gz | samtools sort -@14 -m 3G -o L_RNA_scaffolder.barcoded.bam -
samtools index L_RNA_scaffolder.barcoded.bam
```

```
[M::process] read 1010832 sequences (140000232 bp)...
[M::mem_process_seqs] Processed 1010832 reads in 467.597 CPU sec, 41.840 real sec
[M::process] read 85186 sequences (11798261 bp)...
[M::mem_process_seqs] Processed 1010832 reads in 453.133 CPU sec, 32.656 real sec
[M::mem_process_seqs] Processed 85186 reads in 43.914 CPU sec, 3.325 real sec
[main] Version: 0.7.17-r1188
[main] CMD: bwa mem -t 14 L_RNA_scaffolder.fasta barcoded.fastq.gz
```

[main] Real time: 28825.221 sec; CPU: 346888.296 sec
[bam_sort_core] merging from 84 files and 14 in-memory blocks...

3) Actually running Pilon
Moving forward, we only need to care about the BAM file and the assembly.
It just needs to use a lot of RAM (why we need to run it @qb2). We need to setup an appropriate qsub script for the run:

Running on highmem node

```
#!/bin/bash

#PBS -N 2020-03-14.PilonPolish_scaffolded.pbs
#PBS -l nodes=1:ppn=48
#PBS -l mem=900gb
#PBS -l walltime=12:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

module load bwa/0.7.17
module load samtools/1.10
module load java/8u45
module load pilon/1.23

export _JAVA_OPTIONS="-Xmx900g"

cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/Pilon/bwa-aligned_scaffolded

# run pilon
java -Xmx900G -jar ~/programs/pilon/pilon-1.23.jar --genome L_RNA_scaffolder.fasta --bam L_RNA_scaffolder.barcoded.bam --changes --vcf
--diploid --threads 48 --output L_RNA_scaffolder.polished
```

Ran for about 4 hrs.

```
mkdir slimsute && cd $_
module load python/2.7.15
python ~/SLiMSuite/tools/seqsuite.py summarise batchrun="../*.fasta" basefile=scaffolds dna newlog
```

```
#~~#    00:00:13       # ~~~~~~~~ Sequence Summary for L_RNA_scaffolder ~~~~~~~~~ #
#SUM    00:00:31         Total number of sequences: 7,776
#SUM    00:00:32         Total length of sequences: 1,067,321,776
#SUM    00:00:33         Min. length of sequences: 977
#SUM    00:00:34         Max. length of sequences: 31,181,295
#SUM    00:00:34         Mean length of sequences: 137,258.46
#SUM    00:00:34         Median length of sequences: 2,395
#SUM    00:00:34         N50 length of sequences: 7,118,366
#SUM    00:00:35         L50 count of sequences: 38
#SUM    00:00:35         GC content: 41.82%
#SUM    00:00:35         Gap (N) length: 11,598,876 (1.09%)
#LOAD   00:00:36         Load sequences from ../L_RNA_scaffolder.polished.fasta
#SEQ    00:02:15        7,776 of 7,776 sequences loaded from ../L_RNA_scaffolder.polished.fasta (Format: fas).
#INDEX  00:02:16         Index file ../L_RNA_scaffolder.polished.fasta.index made
#FILT   00:02:17        7,776 of 7,776 sequences retained.
#~~#    00:02:17        # ~~~~ Sequence Summary for L_RNA_scaffolder.polished ~~~~ #
#SUM    00:02:49         Total number of sequences: 7,776
#SUM    00:02:50         Total length of sequences: 1,067,071,200
#SUM    00:02:51         Min. length of sequences: 842
#SUM    00:02:51         Max. length of sequences: 31,169,695
#SUM    00:02:52         Mean length of sequences: 137,226.23
#SUM    00:02:52         Median length of sequences: 2,394
```

```
#SUM    00:02:52    N50 length of sequences: 7,116,007
#SUM    00:02:52    L50 count of sequences: 38
#SUM    00:02:52    GC content: 41.82%
#SUM    00:02:52    Gap (N) length: 11,396,143 (1.07%)
```

```
module load python/3.7.3 blast+/2.2.31 hmmer/3.2.1 augustus/3.3.2 emboss/6.6.0 busco/3.0.2b
export AUGUSTUS_CONFIG_PATH=/srv/scratch/z5188231/programs/augustus
export BUSCO_CONFIG_FILE=/home/z5188231/busco/3.0.2b/config/config.ini
```

```
BUSCOSET=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/data/BUSCO.2018-08-21
```

```
python3 /apps/busco/3.0.2b/scripts/run_BUSCO.py -i ../L_RNA_scaffolder.polished.fasta -o L_RNA_scaffolder.polished.busco -m genome -
l ${BUSCOSET}/aves_odb9/ -c 32 -f
```

```
INFO    Results:
INFO    C:94.5%[S:92.5%,D:2.0%],F:3.4%,M:2.1%,n:4915
INFO    4644 Complete BUSCOs (C)
INFO    4546 Complete and single-copy BUSCOs (S)
INFO    98 Complete and duplicated BUSCOs (D)
INFO    168 Fragmented BUSCOs (F)
INFO    103 Missing BUSCOs (M)
INFO    4915 Total BUSCO groups searched
```