Starling-May18
Projects/Katarina

PDF Version generated by

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

on

Jun 23, 2022 @03:51 PM NZST

Table of Contents

2020-02-03.Scaffold.SSpace.Gapfinisher

Katarina Stuart (z5188231@ad.unsw.edu.au) - Mar 26, 2020, 7:38 PM NZDT

Scaffolding using SSpace & Gapfinisher Gapfilling

NOTE: SSPACE gives different results (minor differences in scaffold number, and hence other measurements) each run.

Can use fastq or fastq for SSPACE but must use fasta for Gapfinisher.

```
#!/bin/bash

#PBS -N SSpacescaffolding.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=128gb
#PBS -l walltime=12:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

module load perl/5.28.0

GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/assembly/Diploidocus/svulgaris-10x-550M-sub80.pri.fasta
READS=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.3_StarlingNanopore/data/basecall/pass/filtered/filtered.fq
OUT_DIR=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/SSPACE/outs

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

perl /apps/sspace/1-1/longread/SSPACE-LongRead.pl -c $GENOME -p $READS -t 16 -k 1 -b $OUT_DIR
```

```
Mon Feb 17 11:37:56 2020: Formatting contig sequences...
Mon Feb 17 11:38:00 2020: Aligning PacBio reads to contigs...
Mon Feb 17 19:04:17 2020: Reading PacBio sequences...
Mon Feb 17 19:04:28 2020: Processing alignment output...
Mon Feb 17 19:04:42 2020: Filtering contig links...
Mon Feb 17 19:04:43 2020: Scaffolding...
Mon Feb 17 19:10:35 2020: Finishing
Scaffolded 18439 contigs into 7855 scaffolds
```

```
module load python/2.7.15
python ~/SLiMSuite/tools/seqsuite.py summarise batchrun="../*.fasta" basefile=scaffolds dna newlog
```

```
#~~#    00:02:42        # ~~~~~~~~~~~~ Sequence Summary for scaffolds ~~~~~~~~~~~~ #
#SUM    00:03:29        Total number of sequences: 7,856
#SUM    00:03:29        Total length of sequences: 1,062,633,441
#SUM    00:03:29        Min. length of sequences: 1,000
#SUM    00:03:29        Max. length of sequences: 30,521,271
#SUM    00:03:29        Mean length of sequences: 135,263.93
#SUM    00:03:29        Median length of sequences: 2,433
#SUM    00:03:29        N50 length of sequences: 7,118,373
#SUM    00:03:29        L50 count of sequences: 39
#SUM    00:03:29        GC content: 41.68%
#SUM    00:03:29        Gap (N) length: 23,897,712 (2.25%)
#SAVE   00:03:29        Table "summarise" saved to "scaffolds.summarise.tdt": 1 entries.
```

```
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 ../scaffolds.fasta -o scaffolds.busco -m genome -l ${BUSCOSET}/aves_odb9/ -c 32 -f
```

```
INFO    Results:
INFO    C:94.4%[S:92.3%,D:2.1%],F:3.5%,M:2.1%,n:4915
INFO    4638 Complete BUSCOs (C)
INFO    4537 Complete and single-copy BUSCOs (S)
INFO    101 Complete and duplicated BUSCOs (D)
INFO    173 Fragmented BUSCOs (F)
INFO    104 Missing BUSCOs (M)
INFO    4915 Total BUSCO groups searched
INFO    BUSCO analysis done. Total running time: 7594.473603963852 seconds
```

Gapfinisher

gapfinisher: https://github.com/kammoji/gapFinisher

PREREQUISITE: A successful run of SSPACE-LongRead so that the directory "inner-scaffold-sequences" is created.

```
#!/bin/bash

#PBS -N 2020-02-24.Gapfinisher.pbs
#PBS -V
#PBS -l nodes=1:ppn=16
#PBS -l mem=24gb
#PBS -l walltime=12:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

module add mcr/2012a
module add gapfinisher/0.1

SSPACE_OUTPUT=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/SSPACE/
READS=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.3_StarlingNanopore/data/basecall/pass/filtered/filtered.fasta
MCR=/apps/mcr/2012a/v717

cd $SSPACE_OUTPUT

gapFinisher -i $SSPACE_OUTPUT -l $READS -m $MCR -t 16
```

Gapfinisher seems to have issues with threading. I have noticed:
- If you run a multi thread job on an interactive node/high mem node, you cannot terminate it. The job keeps running in the background unless you delete the interactive job (and you cannot do this on high mem node).

- I mulithreaded it on high mem node and let it run properly. Output not produced properly (i.e. no scaffolds_gapfilled_FINAL.fasta produced). All threads seem to just end at different times and the final steps with all output happen. Also happened when job submitted to Katana. I am trying to run it now with one thread to see if it works.

```
#!/bin/bash

#PBS -N 2020-02-25.Gapfinisher_v2.pbs
#PBS -V
```

```
#PBS -l nodes=1:ppn=4
#PBS -l mem=24gb
#PBS -l walltime=100:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

module add mcr/2012a
module add gapfinisher/0.1

SSPACE_OUTPUT=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/SSPACE_v2/
READS=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.3_StarlingNanopore/data/basecall/pass/filtered/filtered.fasta
MCR=/apps/mcr/2012a/v717

cd $SSPACE_OUTPUT

gapFinisher -i $SSPACE_OUTPUT -l $READS -m $MCR -t 1
```

With new version of Gapfinisher

https://peerj.com/preprints/3467/

738 days ago - Juhana Kammonen
Dear Users,

There is a problem with the multi-thread mode of the current release of gapFinisher. We are working on a release that will solve the problem. I will link it here when done.

It is safe to use gapFinisher in single-thread mode (option "-t 1") at the moment, although this may be infeasibly slow in some applications.

Please send all your user experiences you think we should know about to juhana.kammonen[at]helsinki.fi

Kind regards,

Juhana I Kammonen

---

169 days ago - Juhana Kammonen

Dear Users,

gapFinisher has just been published in PLoS ONE:

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0216885

The multi-thread mode has finally been fixed. If you installed previous version from GitHub just cd into the folder where you installed gapFinisher (by default named "gapFinisher") and type:

git pull

Fresh installations can be made from:

https://github.com/kammoji/gapFinisher

```
#!/bin/bash

#PBS -N 2020-02-27.V2Gapfinisher.pbs
#PBS -V
#PBS -l nodes=1:ppn=16
```

```
#PBS -l mem=16gb
#PBS -l walltime=12:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

module add mcr/2012a
PATH=$PATH:/home/z5188231/programs/gapFinisher-master/

SSPACE_OUTPUT=/srv/scratch/z5188231/KStuart.Starling-
Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/SSPACE_GapfinisherV2/
READS=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.3_StarlingNanopore/data/basecall/pass/filtered/filtered.fasta
MCR=/apps/mcr/2012a/v717

cd $SSPACE_OUTPUT

gapFinisher -i $SSPACE_OUTPUT -l $READS -m $MCR -t 16
```

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

```
#BATCH  00:00:00       Batch summarising 2 input files
#LOAD   00:00:00       Load sequences from ../scaffolds.fasta
#SEQ    00:00:00       7,856 of 7,856 sequences loaded from ../scaffolds.fasta.index (Format: index).
#FILT   00:00:00       7,856 of 7,856 sequences retained.
#~~#    00:00:00       # ~~~~~~~~~~~~ Sequence Summary for scaffolds ~~~~~~~~~~~~ #
#SUM    00:00:49       Total number of sequences: 7,856
#SUM    00:00:49       Total length of sequences: 1,062,633,441
#SUM    00:00:49       Min. length of sequences: 1,000
#SUM    00:00:49       Max. length of sequences: 30,521,271
#SUM    00:00:49       Mean length of sequences: 135,263.93
#SUM    00:00:49       Median length of sequences: 2,433
#SUM    00:00:49       N50 length of sequences: 7,118,373
#SUM    00:00:49       L50 count of sequences: 39
#SUM    00:00:49       GC content: 41.68%
#SUM    00:00:49       Gap (N) length: 23,897,712 (2.25%)
#LOAD   00:00:49       Load sequences from ../scaffolds_gapfilled_FINAL.fasta
#SEQ    00:03:29       7,856 of 7,856 sequences loaded from ../scaffolds_gapfilled_FINAL.fasta (Format: fas).
#INDEX  00:03:29       Index file ../scaffolds_gapfilled_FINAL.fasta.index made
#FILT   00:03:29       7,856 of 7,856 sequences retained.
#~~#    00:03:29       # ~~~~ Sequence Summary for scaffolds_gapfilled_FINAL ~~~~ #
#SUM    00:04:17       Total number of sequences: 7,856
#SUM    00:04:17       Total length of sequences: 1,067,313,776
#SUM    00:04:17       Min. length of sequences: 977
#SUM    00:04:17       Max. length of sequences: 30,547,435
#SUM    00:04:17       Mean length of sequences: 135,859.70
#SUM    00:04:17       Median length of sequences: 2,433
#SUM    00:04:17       N50 length of sequences: 6,652,296
#SUM    00:04:17       L50 count of sequences: 40
#SUM    00:04:17       GC content: 41.82%
#SUM    00:04:17       Gap (N) length: 11,590,876 (1.09%)
#SAVE   00:04:17       Table "summarise" saved to "scaffolds.summarise.tdt": 2 entries.
```

The N50 has decreased and L50 has increased because genome size has increased I suspect.

```
#!/bin/bash

#PBS -N 2020-02-28.BUSCO_Gapfinisher.pbs
```

```
#PBS -V
#PBS -l nodes=1:ppn=16
#PBS -l mem=16gb
#PBS -l walltime=12:00:00
#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/SSPACE_GapfinisherV2/

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 ../scaffolds.fasta -o scaffolds.busco -m genome -l ${BUSCOSET}/aves_odb9/ -c 32 -f
```