Starling-May18

Projects/Katarina Stuart/KStuart.Starling-Aug18/Sv3_Genome/Assembly/2020-02-02.Diploidocus

PDF Version generated by

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

on

Jun 23, 2022 @03:50 PM NZST

Table of Contents

2020-02-02.Diploidocus

Katarina Stuart (z5188231@ad.unsw.edu.au) - Apr 28, 2020, 3:08 PM NZST

Diploidocus

splits a pseudodiploid assembly into primary and alternative scaffolds: https://github.com/slimsuite/diploidocus

https://slimsuite.github.io/diploidocus/

```
cp  /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/assembly/svulgaris-10x-550M-sub80/outs/fasta/*  /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/Diploidocus

gzip svulgaris-10x-550M-sub80.1.fasta svulgaris-10x-550M-sub80.2.fasta

module load python/2.7.15

zcat svulgaris-10x-550M-sub80.1.fasta | sed 's/>/>phap_STUVU__SVU10XV2PHAP/g' > svu10xv2.pseudodip.fasta
zcat svulgaris-10x-550M-sub80.2.fasta | sed 's/>/>ahap_STUVU__SVU10XV2AHAP/g' >> svu10xv2.pseudodip.fasta
```

```
module load minimap2/2.17

DIR=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/assembly/svulgaris-10x-550M-sub80/outs/fasta
BASEFILE=svulgaris-10x-550M-sub80

python /home/z5188231/programs/diploidocus-master/code/diploidocus.py runmode=diphapnr basefile=svulgaris-10x-550M-sub80 genomesize=1040824271 summarise=T seqin=svu10xv2.pseudodip.fasta seqout=$BASEFILE.nr.fasta

Used 28 forks
```

```
#~~#    04:27:40       # ~~~~ Sequence Summary for svulgaris-10x-550M-sub80.dipnr ~~~~ #
#SUM    04:28:15       Total number of sequences: 19,346
#SUM    04:28:15       Total length of sequences: 1,887,443,050
#SUM    04:28:15       Min. length of sequences: 1,000
#SUM    04:28:15       Max. length of sequences: 12,884,419
#SUM    04:28:15       Mean length of sequences: 97,562.44
#SUM    04:28:15       Median length of sequences: 2,307
#SUM    04:28:15       N50 length of sequences: 1,907,593
#SUM    04:28:15       L50 count of sequences: 245
#SUM    04:28:15       NG50 length of sequences (1.04 Gb): 3,614,482
#SUM    04:28:15       LG50 count of sequences (1.04 Gb): 86
#SUM    04:28:15       GC content: 41.28%
#SUM    04:28:15       Gap (N) length: 11,582,260 (0.61%)
#SEQ    04:28:19       18,439 of 18,439 sequences loaded from svulgaris-10x-550M-sub80.pri.fasta (Format: fas).
#INDEX  04:28:19        Index file svulgaris-10x-550M-sub80.pri.fasta.index made
#FILT   04:28:19       18,439 of 18,439 sequences retained.
#BAK    04:28:42       .nr.fasta backed up as .nr.fasta.bak
#OUT    04:28:49       18,439 Sequences output overwriting .nr.fasta
#~~#    04:28:49       # ~~~~ Sequence Summary for svulgaris-10x-550M-sub80.pri ~~~~ #
#SUM    04:29:09       Total number of sequences: 18,439
#SUM    04:29:09       Total length of sequences: 1,040,106,492
#SUM    04:29:09       Min. length of sequences: 1,000
#SUM    04:29:09       Max. length of sequences: 12,884,419
```

```
#SUM    04:29:09    Mean length of sequences: 56,407.97
#SUM    04:29:09    Median length of sequences: 2,153
#SUM    04:29:09    N50 length of sequences: 1,764,435
#SUM    04:29:09    L50 count of sequences: 146
#SUM    04:29:09    NG50 length of sequences (1.04 Gb): 1,756,637
#SUM    04:29:09    LG50 count of sequences (1.04 Gb): 147
#SUM    04:29:09    GC content: 41.64%
#SUM    04:29:09    Gap (N) length: 6,835,790 (0.66%)
#SEQ    04:29:12    907 of 907 sequences loaded from svulgaris-10x-550M-sub80.alt.fasta (Format: fas).
#INDEX  04:29:12    Index file svulgaris-10x-550M-sub80.alt.fasta.index made
#FILT   04:29:12    907 of 907 sequences retained.
#OUT    04:29:28    907 Sequences output overwriting .nr.fasta
#~~#    04:29:28    # ~~~~ Sequence Summary for svulgaris-10x-550M-sub80.alt ~~~~ #
#SUM    04:29:44    Total number of sequences: 907
#SUM    04:29:44    Total length of sequences: 847,336,558
#SUM    04:29:44    Min. length of sequences: 1,620
#SUM    04:29:44    Max. length of sequences: 12,880,535
#SUM    04:29:44    Mean length of sequences: 934,218.92
#SUM    04:29:44    Median length of sequences: 450,476
#SUM    04:29:44    N50 length of sequences: 2,203,301
#SUM    04:29:44    L50 count of sequences: 102
#SUM    04:29:44    NG50 length of sequences (1.04 Gb): 1,623,183
#SUM    04:29:44    LG50 count of sequences (1.04 Gb): 153
#SUM    04:29:44    GC content: 40.83%
#SUM    04:29:44    Gap (N) length: 4,746,470 (0.56%)
#SAVE   04:29:48    Table "summarise" saved to "svulgaris-10x-550M-sub80.summarise.tdt": 6 entries.
#WARN   04:29:48    2 error messages! Check log for details.
```

RICH:
The alternative assembly will have a bigger N50 as it is smaller, but a smaller NG50. The diploid N(G)50s are inflated by having two copies of all the big scaffolds.

```
#!/bin/bash

#PBS -N 2020-03-04.BUSCO_Diploidocus.pbs
#PBS -V
#PBS -l nodes=1:ppn=40
#PBS -l mem=56gb
#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/assembly/Diploidocus

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 ./svulgaris-10x-550M-sub80.pri.fasta -o svulgaris-10x-550M-sub80.pri -m genome -
l ${BUSCOSET}/aves_odb9/ -c 32 -f
```