Starling-May18
Projects/Katarina

PDF Version generated by

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

on

Jun 23, 2022 @03:49 PM NZST

Table of Contents

2019-12-31.Supernova.BarcodeSubsampling

Katarina Stuart (z5188231@ad.unsw.edu.au) - Sep 11, 2020, 5:38 PM NZST

Supernova assembly with barcode subsampling

Subsampling using Supernova directly

Documentation available at: https://dnatech.genomecenter.ucdavis.edu/wp-content/uploads/2016/01/10X-Genome-Assembly-CG000100__Rev_A_Technical_Note_BCSubsampling_Supernova.pdf

For genomes 0.1 to 1.6 Gb, load less and sequence deeper.

- Load 0.625 ng for any genome in this smaller range. (Do not go lower as this could result in a lowcomplexity library.)
- Sequence deeply: 400-600 M reads for any genome in this smaller range.
- Then, use only a fraction of the barcodes by barcode subsampling (see below). This lowers read coverage of the genome to the optimal range for Supernova, 38-56x, while leaving read depth per molecule unchanged.

Assembly summaries:

- Version 1: 450M, 0.75 barcodes. scaffold N50 = 950 Kb. Lower quality, has been deleted (can be rerun if needed)
- Version 2: 550M, 0.8 barcodes scaffold N50 = 1.84 Mb
- Version 3: 550M, 0.9 barcodes scaffold N50 = 1.76 Mb

VERSION 1:

```
#!/bin/bash

#PBS -N 2019-12-31.Supernova450M-sub75.pbs
#PBS -l nodes=1:ppn=44
#PBS -l mem=350gb
#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/Sv3.2_Starling10x/assembly

module add supernova/2.1.1
```

```
FASTQ=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3.2_Starling10x/data/fastq/
SAMPLE=SV01
```

```
supernova run --id svulgaris-10x-450M-sub75 --fastqs $FASTQ --sample $SAMPLE --description sturnus_vulgaris_10X-450M-sub75 --localcores 44 --localmem 350 --bcfrac=0.75 --maxreads=450000000
```

resources_used.walltime=46:32:20

SUMMARY

- Sun Jan 05 08:59:56 2020
- [svulgaris-10x-450M-sub75] sturnus_vulgaris_10X-450M-sub75
- software release = 2.1.1(6bb16452a)
- likely sequencers = HiSeq X
- assembly checksum = 6,524,032,093,095,280,585

INPUT
- 450.02 M = READS = number of reads; ideal 800M-1200M for human
- 138.50 b = MEAN READ LEN = mean read length after trimming; ideal 140
- 56.21 x = RAW COV = raw coverage; ideal ~56

- 41.19 x = EFFECTIVE COV = effective read coverage; ideal ~42 for raw 56x
- 76.02 % = READ TWO Q30 = fraction of Q30 bases in read 2; ideal 75-85
- 434.00 b = MEDIAN INSERT = median insert size; ideal 350-400
- 82.51 % = PROPER PAIRS = fraction of proper read pairs; ideal >= 75
- 0.75 = BARCODE FRACTION = fraction of barcodes used; between 0 and 1
- 1.20 Gb = EST GENOME SIZE = estimated genome size
- 5.52 % = REPETITIVE FRAC = genome repetitivity index
- 0.04 % = HIGH AT FRACTION = high AT index
- 41.51 % = ASSEMBLY GC CONTENT = GC content of assembly
- 0.56 % = DINUCLEOTIDE FRACTION = dinucleotide content
- 14.06 Kb = MOLECULE LEN = weighted mean molecule size; ideal 50-100
- 33.65 = P10 = molecule count extending 10 kb on both sides
- 235.00 b = HETDIST = mean distance between heterozygous SNPs
- 7.18 % = UNBAR = fraction of reads that are not barcoded
- 450.00 = BARCODE N50 = N50 reads per barcode
- 13.47 % = DUPS = fraction of reads that are duplicates
- 54.55 % = PHASED = nonduplicate and phased reads; ideal 45-50

OUTPUT

- 4.13 K = LONG SCAFFOLDS = number of scaffolds >= 10 kb
- 33.19 Kb = EDGE N50 = N50 edge size
- 126.98 Kb = CONTIG N50 = N50 contig size
- 743.53 Kb = PHASEBLOCK N50 = N50 phase block size
- 925.75 Kb = SCAFFOLD N50 = N50 scaffold size
- 7.89 % = MISSING 10KB = % of base assembly missing from scaffolds >= 10 kb
- 981.92 Mb = ASSEMBLY SIZE = assembly size (only scaffolds >= 10 kb)

Version 2

```
#!/bin/bash

#PBS -N 2019-12-31.Supernova550M-sub80.pbs
#PBS -l nodes=1:ppn=44
#PBS -l mem=350gb
#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/Sv3.2_Starling10x/assembly

module add supernova/2.1.1
```

```
FASTQ=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3.2_Starling10x/data/fastq/
SAMPLE=SV01
```

```
supernova run --id svulgaris-10x-550M-sub80 --fastqs $FASTQ --sample $SAMPLE --description sturnus_vulgaris_10X-550M-sub80 --localcores 44 --localmem 350 --bcfrac=0.8 --maxreads=550000000
```

resources_used.walltime=47:54:04

SUMMARY

- Fri Jan 03 10:27:27 2020
- [svulgaris-10x-550M-sub80] sturnus_vulgaris_10X-550M-sub80
- software release = 2.1.1(6bb16452a)
- likely sequencers = HiSeq X
- assembly checksum = -5,991,739,180,408,811,643

```
INPUT
-  550.01 M  = READS                = number of reads; ideal 800M-1200M for human
-  138.50  b = MEAN READ LEN        = mean read length after trimming; ideal 140
-   69.16  x = RAW COV              = raw coverage; ideal ~56
-   50.19  x = EFFECTIVE COV        = effective read coverage; ideal ~42 for raw 56x
-   76.03  % = READ TWO Q30         = fraction of Q30 bases in read 2; ideal 75-85
-  434.00  b = MEDIAN INSERT        = median insert size; ideal 350-400
-   82.58  % = PROPER PAIRS         = fraction of proper read pairs; ideal >= 75
-    0.80    = BARCODE FRACTION     = fraction of barcodes used; between 0 and 1
-    1.19 Gb = EST GENOME SIZE      = estimated genome size
-    5.37  % = REPETITIVE FRAC      = genome repetitivity index
-    0.05  % = HIGH AT FRACTION     = high AT index
-   41.58  % = ASSEMBLY GC CONTENT  = GC content of assembly
-    0.56  % = DINUCLEOTIDE FRACTION = dinucleotide content
-   14.06 Kb = MOLECULE LEN         = weighted mean molecule size; ideal 50-100
-   42.26    = P10                 = molecule count extending 10 kb on both sides
-  229.00  b = HETDIST              = mean distance between heterozygous SNPs
-    7.15  % = UNBAR                = fraction of reads that are not barcoded
-  516.00    = BARCODE N50          = N50 reads per barcode
-   14.38  % = DUPS                 = fraction of reads that are duplicates
-   55.54  % = PHASED               = nonduplicate and phased reads; ideal 45-50
-------------------------------------------------------------------------------
OUTPUT
-    2.65 K  = LONG SCAFFOLDS       = number of scaffolds >= 10 kb
-   36.64 Kb = EDGE N50             = N50 edge size
-  142.28 Kb = CONTIG N50           = N50 contig size
-    1.12 Mb = PHASEBLOCK N50       = N50 phase block size
-    1.84 Mb = SCAFFOLD N50         = N50 scaffold size
-    6.58  % = MISSING 10KB         = % of base assembly missing from scaffolds >= 10 kb
-  992.02 Mb = ASSEMBLY SIZE        = assembly size (only scaffolds >= 10 kb)
-------------------------------------------------------------------------------
```

```
module load python/2.7.15
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3.2_Starling10x/assembly/svulgaris-10x-550M-sub80/outs/
mkdir fasta
supernova mkoutput --asmdir=assembly --outprefix=fasta/svulgaris-10x-550M-sub80 --style=pseudohap2
cd fasta
gunzip *.gz
python ~/SLiMSuite/tools/seqsuite.py summarise batchrun="*.fasta" basefile=svulgaris-10x-550M-sub80 dna newlog
```

```
#~~#    #~+~~+~#       #~~#
#LOG    00:00:00       Activity Log for SeqSuite V1.23.0: Fri Jan  3 15:46:22 2020
#DIR    00:00:00       Run from directory: /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3.2_Starling10x/assembly/svulgaris-10x-550M-
sub80/outs/fasta
#ARG    00:00:00       Commandline arguments: summarise batchrun=*.fasta basefile=svulgaris-10x-550M-sub80 dna newlog
#CMD    00:00:00       Full Command List: log=seqsuite.log summarise batchrun=*.fasta basefile=svulgaris-10x-550M-sub80 dna newlog
#VIO    00:00:00       Verbosity: 1; Interactivity: 0.
#BASE   00:00:00       svulgaris-10x-550M-sub80
#BATCH  00:00:00       Batch summarising 2 input files
#SEQ    00:01:59       18,778 of 18,778 sequences loaded from svulgaris-10x-550M-sub80.1.fasta (Format: fas).
#INDEX  00:01:59       Index file svulgaris-10x-550M-sub80.1.fasta.index made
#FILT   00:01:59       18,778 of 18,778 sequences retained.
#~~#    00:01:59       # ~~~~ Sequence Summary for svulgaris-10x-550M-sub80.1 ~~~~ #
#SUM    00:02:33       Total number of sequences: 18,778
#SUM    00:02:33       Total length of sequences: 1,040,824,271
#SUM    00:02:33       Min. length of sequences: 1,000
#SUM    00:02:33       Max. length of sequences: 12,884,419
#SUM    00:02:33       Mean length of sequences: 55,427.86
#SUM    00:02:33       Median length of sequences: 2,121
#SUM    00:02:33       N50 length of sequences: 1,756,637
#SUM    00:02:33       L50 count of sequences: 147
#SUM    00:02:33       GC content: 41.64%
#SUM    00:02:33       Gap (N) length: 6,838,110 (0.66%)
```

```
#SEQ    00:04:40        18,778 of 18,778 sequences loaded from svulgaris-10x-550M-sub80.2.fasta (Format: fas).
#INDEX  00:04:40         Index file svulgaris-10x-550M-sub80.2.fasta.index made
#FILT   00:04:40        18,778 of 18,778 sequences retained.
#~~#    00:04:40        # ~~~~ Sequence Summary for svulgaris-10x-550M-sub80.2 ~~~~ #
#SUM    00:05:14         Total number of sequences: 18,778
#SUM    00:05:14         Total length of sequences: 1,040,787,847
#SUM    00:05:14         Min. length of sequences: 1,000
#SUM    00:05:14         Max. length of sequences: 12,880,535
#SUM    00:05:14         Mean length of sequences: 55,425.92
#SUM    00:05:14         Median length of sequences: 2,121
#SUM    00:05:14         N50 length of sequences: 1,756,637
#SUM    00:05:14         L50 count of sequences: 147
#SUM    00:05:14         GC content: 41.64%
#SUM    00:05:14         Gap (N) length: 6,837,230 (0.66%)
#SAVE   00:05:14         Table "summarise" saved to "svulgaris-10x-550M-sub80.summarise.tdt": 2 entries.
#LOG    00:05:14        SeqSuite V1.23.0 End: Fri Jan  3 15:51:36 2020
```

```
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/Sv3.4_GenomeAnnotation/data/BUSCO.2018-08-21
```

```
python3 /apps/busco/3.0.2b/scripts/run_BUSCO.py -i svulgaris-10x-550M-sub80.1.fasta -o svulgaris-10x-550M-sub80.1.busco -m genome -
l ${BUSCOSET}/aves_odb9/ -c 32 -f
```

```
# BUSCO was run in mode: genome

    C:92.8%[S:90.9%,D:1.9%],F:4.5%,M:2.7%,n:4915

    4565    Complete BUSCOs (C)
    4470    Complete and single-copy BUSCOs (S)
    95      Complete and duplicated BUSCOs (D)
    219     Fragmented BUSCOs (F)
    131     Missing BUSCOs (M)
    4915    Total BUSCO groups searched

INFO    BUSCO analysis done. Total running time: 8892.91138100624 seconds
INFO    Results written in /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3.2_Starling10x/assembly/svulgaris-10x-550M-
sub80/outs/fasta/run_svulgaris-10x-550M-sub80.1.busco/
```

Busco output folder deleted to save space.

Version 3

```
#!/bin/bash

#PBS -N 2019-12-31.Supernova550M-sub90.pbs
#PBS -l nodes=1:ppn=44
#PBS -l mem=350gb
#PBS -l walltime=72:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3.2_Starling10x/assembly

module add supernova/2.1.1
```

```
FASTQ=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3.2_Starling10x/data/fastq/
SAMPLE=SV01
```

```
supernova run --id svulgaris-10x-550M-sub90 --fastqs $FASTQ --sample $SAMPLE --description sturnus_vulgaris_10X-550M-sub90 --localcores 44 --localmem 350 --bcfrac=0.9 --maxreads=550000000
```

```
SUMMARY
-------------------------------------------------------------------------------
- Fri Jan 10 03:10:14 2020
- [svulgaris-10x-550M-sub90]  sturnus_vulgaris_10X-550M-sub90
- software release = 2.1.1(6bb16452a)
- likely sequencers = HiSeq X
- assembly checksum = 3,435,794,717,182,712,305
-------------------------------------------------------------------------------
INPUT
-  550.02 M  = READS                = number of reads; ideal 800M-1200M for human
-  138.50  b = MEAN READ LEN        = mean read length after trimming; ideal 140
-   69.09  x = RAW COV              = raw coverage; ideal ~56
-   50.45  x = EFFECTIVE COV        = effective read coverage; ideal ~42 for raw 56x
-   76.04  % = READ TWO Q30         = fraction of Q30 bases in read 2; ideal 75-85
-  434.00  b = MEDIAN INSERT        = median insert size; ideal 350-400
-   82.58  % = PROPER PAIRS         = fraction of proper read pairs; ideal >= 75
-    0.90    = BARCODE FRACTION     = fraction of barcodes used; between 0 and 1
-    1.19 Gb = EST GENOME SIZE      = estimated genome size
-    5.35  % = REPETITIVE FRAC      = genome repetitivity index
-    0.05  % = HIGH AT FRACTION     = high AT index
-   41.57  % = ASSEMBLY GC CONTENT  = GC content of assembly
-    0.56  % = DINUCLEOTIDE FRACTION = dinucleotide content
-   13.92 Kb = MOLECULE LEN         = weighted mean molecule size; ideal 50-100
-   43.32    = P10                 = molecule count extending 10 kb on both sides
-  222.00  b = HETDIST              = mean distance between heterozygous SNPs
-    7.13  % = UNBAR                = fraction of reads that are not barcoded
-  460.00    = BARCODE N50          = N50 reads per barcode
-   13.86  % = DUPS                 = fraction of reads that are duplicates
-   55.57  % = PHASED               = nonduplicate and phased reads; ideal 45-50
-------------------------------------------------------------------------------
OUTPUT
-    2.90 K  = LONG SCAFFOLDS       = number of scaffolds >= 10 kb
-   36.42 Kb = EDGE N50             = N50 edge size
-  138.42 Kb = CONTIG N50           = N50 contig size
-    1.05 Mb = PHASEBLOCK N50       = N50 phase block size
-    1.76 Mb = SCAFFOLD N50         = N50 scaffold size
-    6.73  % = MISSING 10KB         = % of base assembly missing from scaffolds >= 10 kb
-  989.07 Mb = ASSEMBLY SIZE        = assembly size (only scaffolds >= 10 kb)
-------------------------------------------------------------------------------
```

Proc10xG

Splitting data for multiple assemblies. I use these scripts to parse my data into different assemblies based on barcode. Let me know if you run into any issues and I can assist. The documentation can be subpar (or at least was when I ran them for the first time)
https://github.com/ucdavis-bioinformatics/proc10xG

May not attempt this