Starling-May18
Projects/Katarina Stuart/KStuart.Starling-Aug18/Sv3_Genome/Assembly/2020-03-24.DiplodocusCleanup

PDF Version generated by

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

on

Jun 23, 2022 @03:59 PM NZST

Table of Contents

2020-03-24.DiplodocusCleanup

Katarina Stuart (z5188231@ad.unsw.edu.au) - Jun 23, 2022, 3:59 PM NZST

Diploidocus cleanup

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

https://github.com/slimsuite/diploidocus

3.2.1 Dependencies: Diploidocus needs the following programs installed for full functionality

```
module load bbmap blast+ kat minimap2 purge_haplotigs samtools java/8u45 python/3.7.3
```

3.2.2 Input and assembly processing

The main inputs for Diploidocus rating and filtering are:

seqin=FILE : Input sequence assembly to tidy [Required].
screendb=FILE : File of vectors/contaminants to screen out using blastn and VecScreen rules [Optional].
reads=FILELIST: List of fasta/fastq files containing long reads. Wildcard allowed. Can be gzipped. For a single run (not cycling), a BAM file can be supplied instead with bam=FILE. (This will be preferentially used if found, and defaults to $BASEFILE.bam.) Read types (pb/ont) for each file are set with readtype=LIST, which will be cycled if shorter (default=ont). Optionally, the pre-calculated total read length can be provided with readbp=INT and/or the pre-calculated (haploid) genome size can be provided with genomesize=INT.
busco=TSVFILE : BUSCO full table [full_table_$BASEFILE.busco.tsv] used for calculating single copy ("diploid") read depth. This can be over-ridden by setting scdepth=INT.
kmerreads=FILELIST : File of high quality (i.e. short or error-corrected) reads for KAT kmer analysis [Optional]

If a BAM file is not provided/found, Diploidocus will use minimap2 to generate a BAM file of reads=FILELIST data mapped onto the seqin=FILE assembly. Each read file is mapped separately (--secondary=no -L -ax map-ont or --secondary=no -L -ax map-pb) and converted into a sorted BAM files, before merging the BAM files with samtools and indexing the combined file.

Diploidocus will re-use files where they already exist, providing the downstream files are newer than the upstream files. (If files have been copied and lost their datestamp information, switching ignoredate=T will re-use files regardless.) Setting force=T should force regeneration of files even if they exist.

Version 1: Using Nala rating

3.2.5.4 Nala rating

The `purgemode=nala` rating scheme was used for the Nala German Shepherd Dog genome assembly, and features a simplified set of ratings:

- `CONTAMINATION` = 50%+ identified contamination (`ScreenPerc`)
- `LOWCOV` = Poor median read coverage (`Median_fold < minmedian=INT`)
- `LOWQUAL` = Scaffolds below the sequence length set by `minlen=INT`
- `HPURGE` = Any scaffold with 80%+ bases in the low/haploid coverage bins (haplotigs or assembly artefacts).
- `PRIMARY` = Scaffolds with 20%+ diploid coverage are marked as retention as probable diploids.
- `COLLAPSED` = Scaffolds with <20% diploid coverage and 50%+ high coverage are marked as probable collapsed repeats.
- `JUNK/HAPLOTIG/KEEP` = Remaining Scaffolds are given the PurgeHaplotigs rating (over 80% low/high coverage will be filtered as a probable artefact)

seems that python 2.7 is not installed I get syntax error warnings.

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

module purge
module load python/3.7.3 kat/2.4.2 perl/5.28.0 bedtools/2.27.1 R/3.5.3 samtools/1.10 purge_haplotigs/20190612 java/8u231-jre bbmap/38.51 minimap2/2.17 blast+/2.9.0 python/2.7.15
```

Error at minimap2 stage when running the original genome fasta file, caused by | characters in the names

To fix, run `sed -i 's/|/_/g'` on your input file prior to running and that should fix it. `sed -i` does an in place replacement, if this would cause issues for other things, use regular sed into a new file.

```
cp /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/Pilon/bwa-aligned_scaffolded/L_RNA_scaffolder.polished.fasta .
sed -i 's/|/_/g' L_RNA_scaffolder.polished.fasta
```

Now run Diplodocus tidy

```
GENOME=L_RNA_scaffolder.polished.fasta
BUSCO=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/Pilon/bwa-aligned_scaffolded/slimsute/run_L_RNA_scaffolder.polished.busco/full_table_L_RNA_scaffolder.polished.busco.tsv
READS=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.3_StarlingNanopore/data/basecall/pass/filtered/filtered.fq
```

```
KMERREADS="/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/data/fastq/SV01_S1_L006_R*_001.fastq"
SCREENDB=/home/z5188231/programs/diploidocus-master/data/vecscreendb.fasta
PREFIX=L_RNA_scaffolder.polished.tidy
PPN=1

python /home/z5188231/programs/diploidocus-master/code/diploidocus.py -seqin $GENOME -basefile $PREFIX -busco $BUSCO -reads $READS
kmerreads=\"$KMERREADS\" -forks $PPN 10xtrim=F -screendb $SCREENDB -purgemode nala -runmode dipcycle -pretrim T -veccheck T
```

Errors out at Kat step. Intermediate purge file results below:

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

```
#~~#    00:00:01       # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.1.purge.artefacts ~~~~ #
#SUM    00:00:02       Total number of sequences: 258
#SUM    00:00:02       Total length of sequences: 3,871,806
#SUM    00:00:02       Min. length of sequences: 1,002
#SUM    00:00:02       Max. length of sequences: 63,616
#SUM    00:00:02       Mean length of sequences: 15,007.00
#SUM    00:00:02       Median length of sequences: 12,344
#SUM    00:00:02       N50 length of sequences: 24,279
#SUM    00:00:02       L50 count of sequences: 55
#SUM    00:00:02       GC content: 49.00%
#SUM    00:00:02       Gap (N) length: 146,385 (3.78%)
#LOAD   00:00:02       Load sequences from ../L_RNA_scaffolder.polished.tidy.1.purge.fasta
#SEQ    00:07:32       6,221 of 6,221 sequences loaded from ../L_RNA_scaffolder.polished.tidy.1.purge.fasta (Format: fas).
#INDEX  00:07:32       Index file ../L_RNA_scaffolder.polished.tidy.1.purge.fasta.index made
#FILT   00:07:33       6,221 of 6,221 sequences retained.
#~~#    00:07:33       # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.1.purge ~~~~ #
#SUM    00:08:17       Total number of sequences: 6,221
#SUM    00:08:17       Total length of sequences: 1,047,747,551
#SUM    00:08:17       Min. length of sequences: 917
#SUM    00:08:17       Max. length of sequences: 31,169,695
#SUM    00:08:17       Mean length of sequences: 168,421.08
#SUM    00:08:17       Median length of sequences: 2,198
#SUM    00:08:17       N50 length of sequences: 7,615,694
#SUM    00:08:17       L50 count of sequences: 37
#SUM    00:08:17       GC content: 41.73%
#SUM    00:08:17       Gap (N) length: 11,158,567 (1.07%)
#LOAD   00:08:17       Load sequences from ../L_RNA_scaffolder.polished.tidy.1.purge.haplotigs.fasta
#SEQ    00:08:24       1,297 of 1,297 sequences loaded from ../L_RNA_scaffolder.polished.tidy.1.purge.haplotigs.fasta (Format: fas).
#INDEX  00:08:24       Index file ../L_RNA_scaffolder.polished.tidy.1.purge.haplotigs.fasta.index made
#FILT   00:08:24       1,297 of 1,297 sequences retained.
#~~#    00:08:24       # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.1.purge.haplotigs ~~~~ #
#SUM    00:08:25       Total number of sequences: 1,297
#SUM    00:08:25       Total length of sequences: 15,451,843
#SUM    00:08:25       Min. length of sequences: 842
#SUM    00:08:25       Max. length of sequences: 3,321,351
#SUM    00:08:25       Mean length of sequences: 11,913.53
#SUM    00:08:25       Median length of sequences: 3,044
#SUM    00:08:25       N50 length of sequences: 37,941
#SUM    00:08:25       L50 count of sequences: 63
#SUM    00:08:25       GC content: 46.34%
#SUM    00:08:25       Gap (N) length: 91,191 (0.59%)
#SAVE   00:08:25       Table "summarise" saved to "scaffolds.summarise.tdt": 3 entries.
```

Total number of sequences: 7,776 -> 6,221
Total length of sequences: 1,067,071,200 -> 1,047,747,551
Mean length of sequences: 137,226.23 -> 168,421.08
Median length of sequences: 2,394 -> 2,198
N50 length of sequences: 7,116,007 -> 7,615,694

Got the exact same outputs with the below 3 variations of the command:

```
python /home/z5188231/programs/diploidocus-master/code/diploidocus.py -seqin $GENOME -basefile $PREFIX -busco $BUSCO -reads $READS
kmerreads=\"$KMERREADS\" -forks $PPN 10xtrim=F -screendb $SCREENDB -purgemode nala -runmode dipcycle -pretrim T -veccheck T

python /home/z5188231/programs/diploidocus-master/code/diploidocus.py -seqin $GENOME -basefile $PREFIX -busco $BUSCO -reads $READS
kmerreads=\"$KMERREADS\" -forks $PPN 10xtrim=F -screendb $SCREENDB -purgemode nala -runmode dipcycle -pretrim T -veccheck T -minmedian 0 -deptrim 0

python /home/z5188231/programs/diploidocus-master/code/diploidocus.py -seqin $GENOME -basefile $PREFIX -busco $BUSCO -reads $READS
kmerreads=\"$KMERREADS\" -forks $PPN 10xtrim=F -screendb $SCREENDB -purgemode complex -runmode dipcycle -pretrim T -veccheck T
```

- I did not re-run busco on scaffold name edited contig fasta. Rerunning busco now and will then redo some of the above. My commands do not seem to be working at the moment so this may be why?
- Also, trying to get it to run to completion. I believe that it is a memory problem. Have upped memory from 56 -> 124 -> 248gb
- Also the lack of depth in the long reads may be an issue

Working through the errors:

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

GENOME=./L_RNA_scaffolder.polished.fasta
BUSCO=../run_L_RNA_scaffolder.polished.busco/full_table_L_RNA_scaffolder.polished.busco.tsv
READS=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.3_StarlingNanopore/data/basecall/pass/filtered/filtered.fq
KMERREADS="/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/data/fastq/SV01_S1_L006_R*_001.fastq"
SCREENDB=/home/z5188231/programs/diploidocus-master/data/vecscreendb.fasta
PREFIX=L_RNA_scaffolder.polished.tidy

module add python/3.7.3 kat/2.4.2 perl/5.28.0 bedtools/2.27.1 R/3.5.3 samtools/1.10 purge_haplotigs/20190612 java/8u231-jre bbmap/38.51 minimap2/2.17
blast+/2.9.0 python/2.7.15

python /home/z3452659/slimsuitedev/tools/diploidocus.py -seqin $GENOME -runmode dipcycle -basefile $PREFIX -busco $BUSCO -reads
$READS kmerreads=\"$KMERREADS\" 10xtrim=T 10xtrim -forks 16 -screendb $SCREENDB pretrim=T
```

The above stopped here when requesting 56 GB:

```
[30-03-2020 17:09:01]   Contig scaffold6986_size1108_pilon added back to primary assembly
[30-03-2020 17:09:01]

GENERATING OUTPUT

[30-03-2020 17:09:01] Writing contig associations
[30-03-2020 17:09:02] Writing the reassignment table and new assembly files
[30-03-2020 17:09:11]

PURGE HAPLOTIGS HAS COMPLETED SUCCESSFULLY!

#SYSEND 00:26:54        PURGE HAPLOTIGS HAS COMPLETED SUCCESSFULLY!
#CHECK  00:26:54        L_RNA_scaffolder.polished.tidy.1.bam.gencov: Found.
#CHECK  00:26:54        L_RNA_scaffolder.polished.tidy.1.purge.coverage_stats.csv: Found.
#CHECK  00:26:54        L_RNA_scaffolder.polished.tidy.1.purge.reassignments.tsv: Found.
#CHECK  00:26:54        L_RNA_scaffolder.polished.tidy.1.kat-stats.tsv: Not found: will generate.
#CHECK  00:26:54        L_RNA_scaffolder.polished.tidy.1.kat-counts.cvg: Not found: will generate.
#SYS    00:26:54        kat sect -t 16 --5ptrim 16,0 -o L_RNA_scaffolder.polished.tidy.1.kat ./L_RNA_scaffolder.polished.fasta /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/data/fastq/SV01_S1_L006_R1_001.fastq /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/data/fastq/SV01_S1_L006_R2_001.fastq
=>> PBS: job killed: walltime 39639 exceeded limit 39600
```

Runing the error line on its own at 124 gb:

```
kat sect -t 16 --5ptrim 16,0 -o L_RNA_scaffolder.polished.tidy.1.kat ./L_RNA_scaffolder.polished.fasta /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/data/fastq/SV01_S1_L006_R1_001.fastq /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/data/fastq/SV01_S1_L006_R2_001.fastq
```

On 128 GB mem 16 ppn

```
Warning: Specified hash size insufficent - attempting to double hash size...
Warning: Specified hash size insufficent - attempting to double hash size...
Warning: Specified hash size insufficent - attempting to double hash size...
Warning: Specified hash size insufficent - attempting to double hash size...../deps/seqan-library-2.0.0/include/seqan/basic/basic_exception.h:368 FAILED!  (../deps/seqan-library-2.0.0/include/seqan/basic/basic_exception.h:368 FAILED!  (Uncaught exception of type std::runtime_error: Hash full)
Uncaught exception of type std::runtime_error: Hash full)

../deps/seqan-library-2.0.0/include/seqan/basic/basic_exception.h:368../deps/seqan-library-2.0.0/include/seqan/basic/basic_exception.h../deps/seqan-library-2.0.0/include/seqan/basic/basic_exception.hAborted
```

Trying to run again on 248gb mem and 16 ppn.

Slightly altered command as below at 5ptrim flag:

```
kat sect -t 16 --5ptrim 16 -o L_RNA_scaffolder.polished.tidy.1.kat ./L_RNA_scaffolder.polished.fasta /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/data/fastq/SV01_S1_L006_R1_001.fastq /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/data/fastq/SV01_S1_L006_R2_001.fastq
```

success!
done. Time taken: 4262.0s

-then went on to do the second lot of reads, but I stopped the process.

Running Diplodocus + Nala + Extra + 10x set:

```
#!/bin/bash

#PBS -N 2020-03-29.DiplodocusNalaExtra.pbs
#PBS -V
#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
```

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

module purge
module load python/3.7.3 kat/2.4.2 perl/5.28.0 bedtools/2.27.1 R/3.5.3 samtools/1.10 purge_haplotigs/20190612 java/8u231-jre bbmap/38.51 minimap2/2.17
blast+/2.9.0 python/2.7.15
```

```
GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/Diplodocus_tidy_all/L_RNA_scaffolder.polished.fasta
BUSCO=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/Diplodocus_tidy_all/run_L_RNA_scaffolder.polished.busco/full_table_L_RNA_scaffolder.polished.busco.tsv
READS=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.3_StarlingNanopore/data/basecall/pass/filtered/filtered.fq
KMERREADS="/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/data/fastq/SV01_S1_L006_R*_001.fastq"
SCREENDB=/home/z5188231/programs/diploidocus-master/data/vecscreendb.fasta
PREFIX=L_RNA_scaffolder.polished.tidy
PPN=16

python /home/z5188231/programs/diploidocus-master/code/diploidocus.py -seqin $GENOME -basefile $PREFIX -busco $BUSCO -reads $READS
kmerreads=\"$KMERREADS\" -forks $PPN 10xtrim=T -screendb $SCREENDB -purgemode nala -runmode diploidocus -pretrim T -veccheck T -minmedian 0 -deptrim 0
```

```
#~~#    01:44:38       # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.diploidocus ~~~~ #
#SUM    01:44:57        Total number of sequences: 2,442
#SUM    01:44:57        Total length of sequences: 1,034,661,144
#SUM    01:44:57        Min. length of sequences: 1,001
#SUM    01:44:57        Max. length of sequences: 31,169,695
#SUM    01:44:57        Mean length of sequences: 423,694.16
#SUM    01:44:57        Median length of sequences: 14,757
#SUM    01:44:57        N50 length of sequences: 7,824,914
#SUM    01:44:57        L50 count of sequences: 36
#SUM    01:44:57        GC content: 41.58%
#SUM    01:44:57        Gap (N) length: 10,124,290 (0.98%)
```

Running DipCycle + Nala set:

```
#!/bin/bash

#PBS -N 2020-03-30.DipCycleNala.pbs
#PBS -V
#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
```

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

```

```
GENOME=./L_RNA_scaffolder.polished.fasta
BUSCO=../run_L_RNA_scaffolder.polished.busco/full_table_L_RNA_scaffolder.polished.busco.tsv
READS=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.3_StarlingNanopore/data/basecall/pass/filtered/filtered.fq
KMERREADS="/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/data/fastq/SV01_S1_L006_R*_001.fastq"
SCREENDB=/home/z5188231/programs/diploidocus-master/data/vecscreendb.fasta
PREFIX=L_RNA_scaffolder.polished.tidy

module add python/3.7.3 kat/2.4.2 perl/5.28.0 bedtools/2.27.1 R/3.5.3 samtools/1.10 purge_haplotigs/20190612 java/8u231-jre bbmap/38.51 minimap2/2.17
```

```
blast+/2.9.0 python/2.7.15

python /home/z3452659/slimsuitedev/tools/diploidocus.py -seqin $GENOME -runmode dipcycle -purgemode nala -basefile $PREFIX -busco $BUSCO -reads
$READS kmerreads=\"$KMERREADS\" 10xtrim=T 10xtrim -forks 16 -screendb $SCREENDB pretrim=T
```

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

```
#~~#    00:00:00        # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.diploidocus ~~~~ #
#SUM    00:00:18        Total number of sequences: 2,131
#SUM    00:00:18        Total length of sequences: 1,021,623,305
#SUM    00:00:18        Min. length of sequences: 1,001
#SUM    00:00:18        Max. length of sequences: 31,169,695
#SUM    00:00:18        Mean length of sequences: 479,410.28
#SUM    00:00:18        Median length of sequences: 14,671
#SUM    00:00:18        N50 length of sequences: 8,560,996
#SUM    00:00:18        L50 count of sequences: 35
#SUM    00:00:18        GC content: 41.51%
#SUM    00:00:18        Gap (N) length: 8,552,324 (0.84%)
```

Running DipCycle set:

```
#!/bin/bash

#PBS -N 2020-03-30.DipCycle.pbs
#PBS -V
#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
```

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

```

GENOME=./L_RNA_scaffolder.polished.fasta
BUSCO=../run_L_RNA_scaffolder.polished.busco/full_table_L_RNA_scaffolder.polished.busco.tsv
READS=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.3_StarlingNanopore/data/basecall/pass/filtered/filtered.fq
KMERREADS="/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/data/fastq/SV01_S1_L006_R*_001.fastq"
SCREENDB=/home/z5188231/programs/diploidocus-master/data/vecscreendb.fasta
PREFIX=L_RNA_scaffolder.polished.tidy

module add python/3.7.3 kat/2.4.2 perl/5.28.0 bedtools/2.27.1 R/3.5.3 samtools/1.10 purge_haplotigs/20190612 java/8u231-jre bbmap/38.51 minimap2/2.17
blast+/2.9.0 python/2.7.15

python /home/z3452659/slimsuitedev/tools/diploidocus.py -seqin $GENOME -runmode dipcycle -basefile $PREFIX -busco $BUSCO -reads
$READS kmerreads=\"$KMERREADS\" 10xtrim=T 10xtrim -forks 16 -screendb $SCREENDB pretrim=T
```

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

```
#~~#    00:00:00        # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.1.diploidocus ~~~~ #
#SUM    00:00:21        Total number of sequences: 2,170
#SUM    00:00:21        Total length of sequences: 1,020,540,436
#SUM    00:00:21        Min. length of sequences: 1,000
#SUM    00:00:21        Max. length of sequences: 31,169,695
#SUM    00:00:21        Mean length of sequences: 470,295.13
#SUM    00:00:21        Median length of sequences: 13,151
#SUM    00:00:21        N50 length of sequences: 8,560,996
#SUM    00:00:21        L50 count of sequences: 35
#SUM    00:00:21        GC content: 41.51%
#SUM    00:00:21        Gap (N) length: 8,561,859 (0.84%)
#WARN   00:00:21        summarise entry "../L_RNA_scaffolder.polished.tidy.1.diploidocus.fasta" being overwritten
Suppress future "entry_overwrite" warnings? (y/n) [default=Y]:  n
```

```
#LOAD   00:00:28    Load sequences from ../L_RNA_scaffolder.polished.tidy.2.diploidocus.fasta
#SEQ    00:00:28    2,100 of 2,100 sequences loaded from ../L_RNA_scaffolder.polished.tidy.2.diploidocus.fasta.index (Format: index).
#FILT   00:00:28    2,100 of 2,100 sequences retained.
#~~#    00:00:28    # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.2.diploidocus ~~~~ #
#SUM    00:00:49    Total number of sequences: 2,100
#SUM    00:00:49    Total length of sequences: 1,019,569,463
#SUM    00:00:49    Min. length of sequences: 1,000
#SUM    00:00:49    Max. length of sequences: 31,169,695
#SUM    00:00:49    Mean length of sequences: 485,509.27
#SUM    00:00:49    Median length of sequences: 13,423
#SUM    00:00:49    N50 length of sequences: 8,560,996
#SUM    00:00:49    L50 count of sequences: 35
#SUM    00:00:49    GC content: 41.50%
#SUM    00:00:49    Gap (N) length: 8,536,636 (0.84%)
#~~#    00:00:44    # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.3.diploidocus ~~~~ #
#SUM    00:01:06    Total number of sequences: 2,076
#SUM    00:01:06    Total length of sequences: 1,019,201,667
#SUM    00:01:06    Min. length of sequences: 1,000
#SUM    00:01:06    Max. length of sequences: 31,169,695
#SUM    00:01:06    Mean length of sequences: 490,944.93
#SUM    00:01:06    Median length of sequences: 13,423
#SUM    00:01:06    N50 length of sequences: 8,560,996
#SUM    00:01:06    L50 count of sequences: 35
#SUM    00:01:06    GC content: 41.49%
#SUM    00:01:06    Gap (N) length: 8,533,342 (0.84%)
#LOAD   00:01:06    Load sequences from ../L_RNA_scaffolder.polished.tidy.4.diploidocus.fasta
#SEQ    00:01:06    2,061 of 2,061 sequences loaded from ../L_RNA_scaffolder.polished.tidy.4.diploidocus.fasta.index (Format: index).
#FILT   00:01:06    2,061 of 2,061 sequences retained.
#~~#    00:01:06    # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.4.diploidocus ~~~~ #
#SUM    00:01:29    Total number of sequences: 2,061
#SUM    00:01:29    Total length of sequences: 1,018,961,034
#SUM    00:01:29    Min. length of sequences: 1,000
#SUM    00:01:29    Max. length of sequences: 31,169,695
#SUM    00:01:29    Mean length of sequences: 494,401.28
#SUM    00:01:29    Median length of sequences: 13,420
#SUM    00:01:29    N50 length of sequences: 8,560,996
#SUM    00:01:29    L50 count of sequences: 35
#SUM    00:01:29    GC content: 41.49%
#SUM    00:01:29    Gap (N) length: 8,532,322 (0.84%)
#LOAD   00:01:29    Load sequences from ../L_RNA_scaffolder.polished.tidy.5.diploidocus.fasta
#SEQ    00:01:29    2,056 of 2,056 sequences loaded from ../L_RNA_scaffolder.polished.tidy.5.diploidocus.fasta.index (Format: index).
#FILT   00:01:29    2,056 of 2,056 sequences retained.
#~~#    00:01:29    # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.5.diploidocus ~~~~ #
#SUM    00:01:52    Total number of sequences: 2,056
#SUM    00:01:52    Total length of sequences: 1,018,911,271
#SUM    00:01:52    Min. length of sequences: 1,000
#SUM    00:01:52    Max. length of sequences: 31,169,695
#SUM    00:01:52    Mean length of sequences: 495,579.41
#SUM    00:01:52    Median length of sequences: 13,423
#SUM    00:01:52    N50 length of sequences: 8,560,996
#SUM    00:01:52    L50 count of sequences: 35
#SUM    00:01:52    GC content: 41.49%
#SUM    00:01:52    Gap (N) length: 8,532,211 (0.84%)
#LOAD   00:01:52    Load sequences from ../L_RNA_scaffolder.polished.tidy.6.diploidocus.fasta
#SEQ    00:01:52    2,050 of 2,050 sequences loaded from ../L_RNA_scaffolder.polished.tidy.6.diploidocus.fasta.index (Format: index).
#FILT   00:01:52    2,050 of 2,050 sequences retained.
#~~#    00:01:52    # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.6.diploidocus ~~~~ #
#SUM    00:02:15    Total number of sequences: 2,050
#SUM    00:02:15    Total length of sequences: 1,018,827,636
#SUM    00:02:15    Min. length of sequences: 1,000
#SUM    00:02:15    Max. length of sequences: 31,169,695
#SUM    00:02:15    Mean length of sequences: 496,989.09
#SUM    00:02:15    Median length of sequences: 13,416
#SUM    00:02:15    N50 length of sequences: 8,560,996
#SUM    00:02:15    L50 count of sequences: 35
#SUM    00:02:15    GC content: 41.49%
#SUM    00:02:15    Gap (N) length: 8,531,779 (0.84%)
#LOAD   00:02:15    Load sequences from ../L_RNA_scaffolder.polished.tidy.7.diploidocus.fasta
#SEQ    00:02:15    2,045 of 2,045 sequences loaded from ../L_RNA_scaffolder.polished.tidy.7.diploidocus.fasta.index (Format: index).
#FILT   00:02:15    2,045 of 2,045 sequences retained.
#~~#    00:02:15    # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.7.diploidocus ~~~~ #
#SUM    00:02:38    Total number of sequences: 2,045
#SUM    00:02:38    Total length of sequences: 1,018,756,078
#SUM    00:02:38    Min. length of sequences: 1,000
#SUM    00:02:38    Max. length of sequences: 31,169,695
#SUM    00:02:38    Mean length of sequences: 498,169.23
#SUM    00:02:38    Median length of sequences: 13,427
```

```
#SUM    00:02:38        N50 length of sequences: 8,560,996
#SUM    00:02:38        L50 count of sequences: 35
#SUM    00:02:38        GC content: 41.49%
#SUM    00:02:38        Gap (N) length: 8,531,669 (0.84%)
#SAVE   00:02:38        Table "summarise" saved to "scaffolds.summarise.tdt": 7 entries.
#LOG    00:02:38        SeqSuite V1.23.0 End: Wed Apr  1 11:28:50 2020
```

Running DipCycle + Nala set:

```
#!/bin/bash

#PBS -N 2020-04-01.DipCycleNalaExtra.pbs
#PBS -V
#PBS -l nodes=1:ppn=16
#PBS -l mem=248gb
#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/Diplodocus_tidy_all/DipCycyle_Nala_Extra
```

```
GENOME=./L_RNA_scaffolder.polished.fasta
BUSCO=../run_L_RNA_scaffolder.polished.busco/full_table_L_RNA_scaffolder.polished.busco.tsv
READS=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.3_StarlingNanopore/data/basecall/pass/filtered/filtered.fq
KMERREADS="/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/data/fastq/SV01_S1_L006_R*_001.fastq"
SCREENDB=/home/z5188231/programs/diploidocus-master/data/vecscreendb.fasta
PREFIX=L_RNA_scaffolder.polished.tidy

module add python/3.7.3 kat/2.4.2 perl/5.28.0 bedtools/2.27.1 R/3.5.3 samtools/1.10 purge_haplotigs/20190612 java/8u231-jre bbmap/38.51 minimap2/2.17
blast+/2.9.0 python/2.7.15

python /home/z3452659/slimsuitedev/tools/diploidocus.py -seqin $GENOME -runmode dipcycle -purgemode nala -basefile $PREFIX -busco $BUSCO -reads
$READS kmerreads=\"$KMERREADS\" 10xtrim=T 10xtrim -forks 16 -screendb $SCREENDB pretrim=T -minmedian 0 -deptrim 0
```

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

```
#~~#    00:00:00        # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.1.diploidocus ~~~~ #
#SUM    00:00:17        Total number of sequences: 2,442
#SUM    00:00:17        Total length of sequences: 1,034,661,144
#SUM    00:00:17        Min. length of sequences: 1,001
#SUM    00:00:17        Max. length of sequences: 31,169,695
#SUM    00:00:17        Mean length of sequences: 423,694.16
#SUM    00:00:17        Median length of sequences: 14,757
#SUM    00:00:17        N50 length of sequences: 7,824,914
#SUM    00:00:17        L50 count of sequences: 36
#SUM    00:00:17        GC content: 41.58%
#SUM    00:00:17        Gap (N) length: 10,124,290 (0.98%)
#~~#    00:00:18        # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.2.diploidocus ~~~~ #
#SUM    00:00:37        Total number of sequences: 2,410
#SUM    00:00:37        Total length of sequences: 1,034,467,555
#SUM    00:00:37        Min. length of sequences: 1,001
#SUM    00:00:37        Max. length of sequences: 31,169,695
#SUM    00:00:37        Mean length of sequences: 429,239.65
#SUM    00:00:37        Median length of sequences: 14,940
#SUM    00:00:37        N50 length of sequences: 7,824,914
#SUM    00:00:37        L50 count of sequences: 36
#SUM    00:00:37        GC content: 41.58%
#SUM    00:00:37        Gap (N) length: 10,112,424 (0.98%)
#LOAD   00:00:37        Load sequences from ../L_RNA_scaffolder.polished.tidy.3.diploidocus.fasta
#SEQ    00:00:37        2,405 of 2,405 sequences loaded from ../L_RNA_scaffolder.polished.tidy.3.diploidocus.fasta.index (Format: index).
#FILT   00:00:37        2,405 of 2,405 sequences retained.
#~~#    00:00:37        # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.3.diploidocus ~~~~ #
#SUM    00:00:56        Total number of sequences: 2,405
#SUM    00:00:56        Total length of sequences: 1,034,429,581
#SUM    00:00:56        Min. length of sequences: 1,001
#SUM    00:00:56        Max. length of sequences: 31,169,695
#SUM    00:00:56        Mean length of sequences: 430,116.25
#SUM    00:00:56        Median length of sequences: 14,969
```

```
#SUM    00:00:56    N50 length of sequences: 7,824,914
#SUM    00:00:56    L50 count of sequences: 36
#SUM    00:00:56    GC content: 41.58%
#SUM    00:00:56    Gap (N) length: 10,112,120 (0.98%)
#LOAD   00:00:56    Load sequences from ../L_RNA_scaffolder.polished.tidy.4.diploidocus.fasta
#SEQ    00:00:56    2,403 of 2,403 sequences loaded from ../L_RNA_scaffolder.polished.tidy.4.diploidocus.fasta.index (Format: index).
#FILT   00:00:56    2,403 of 2,403 sequences retained.
#~~#    00:00:56    # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.4.diploidocus ~~~~ #
#SUM    00:01:14    Total number of sequences: 2,403
#SUM    00:01:14    Total length of sequences: 1,034,416,462
#SUM    00:01:14    Min. length of sequences: 1,001
#SUM    00:01:14    Max. length of sequences: 31,169,695
#SUM    00:01:14    Mean length of sequences: 430,468.77
#SUM    00:01:14    Median length of sequences: 14,977
#SUM    00:01:14    N50 length of sequences: 7,824,914
#SUM    00:01:14    L50 count of sequences: 36
#SUM    00:01:14    GC content: 41.58%
#SUM    00:01:14    Gap (N) length: 10,111,910 (0.98%)
#LOAD   00:01:15    Load sequences from ../L_RNA_scaffolder.polished.tidy.5.diploidocus.fasta
#SEQ    00:01:15    2,402 of 2,402 sequences loaded from ../L_RNA_scaffolder.polished.tidy.5.diploidocus.fasta.index (Format: index).
#FILT   00:01:15    2,402 of 2,402 sequences retained.
#~~#    00:01:15    # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.5.diploidocus ~~~~ #
#SUM    00:01:33    Total number of sequences: 2,402
#SUM    00:01:33    Total length of sequences: 1,034,414,968
#SUM    00:01:33    Min. length of sequences: 1,001
#SUM    00:01:33    Max. length of sequences: 31,169,695
#SUM    00:01:33    Mean length of sequences: 430,647.36
#SUM    00:01:33    Median length of sequences: 15,001
#SUM    00:01:33    N50 length of sequences: 7,824,914
#SUM    00:01:33    L50 count of sequences: 36
#SUM    00:01:33    GC content: 41.58%
#SUM    00:01:33    Gap (N) length: 10,111,910 (0.98%)
#LOAD   00:01:33    Load sequences from ../L_RNA_scaffolder.polished.tidy.6.diploidocus.fasta
#SEQ    00:01:33    2,402 of 2,402 sequences loaded from ../L_RNA_scaffolder.polished.tidy.6.diploidocus.fasta.index (Format: index).
#FILT   00:01:33    2,402 of 2,402 sequences retained.
#~~#    00:01:33    # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.6.diploidocus ~~~~ #
#SUM    00:01:52    Total number of sequences: 2,402
#SUM    00:01:52    Total length of sequences: 1,034,414,968
#SUM    00:01:52    Min. length of sequences: 1,001
#SUM    00:01:52    Max. length of sequences: 31,169,695
#SUM    00:01:52    Mean length of sequences: 430,647.36
#SUM    00:01:52    Median length of sequences: 15,001
#SUM    00:01:52    N50 length of sequences: 7,824,914
#SUM    00:01:52    L50 count of sequences: 36
#SUM    00:01:52    GC content: 41.58%
#SUM    00:01:52    Gap (N) length: 10,111,910 (0.98%)
```

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

```
#~~#    00:00:00    # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.diploidocus ~~~~ #
#SUM    00:00:18    Total number of sequences: 2,402
#SUM    00:00:18    Total length of sequences: 1,034,414,968
#SUM    00:00:18    Min. length of sequences: 1,001
#SUM    00:00:18    Max. length of sequences: 31,169,695
#SUM    00:00:18    Mean length of sequences: 430,647.36
#SUM    00:00:18    Median length of sequences: 15,001
#SUM    00:00:18    N50 length of sequences: 7,824,914
#SUM    00:00:18    L50 count of sequences: 36
#SUM    00:00:18    GC content: 41.58%
#SUM    00:00:18    Gap (N) length: 10,111,910 (0.98%)
```

```
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.tidy.diploidocus.fasta -o L_RNA_scaffolder.polished.tidy.diploidocus.busco -m genome -l ${BUSCOSET}/aves_odb9/ -c 32 -f
```

```
INFO    Results:
INFO    C:94.4%[S:93.2%,D:1.2%],F:3.3%,M:2.3%,n:4915
INFO    4641 Complete BUSCOs (C)
INFO    4580 Complete and single-copy BUSCOs (S)
INFO    61 Complete and duplicated BUSCOs (D)
INFO    164 Fragmented BUSCOs (F)
INFO    110 Missing BUSCOs (M)
INFO    4915 Total BUSCO groups searched
```

Which ones not mapping

FINAL DECISION DIPLODOCUS CLEANUP STEP (above too strict for depth of long read data):

Running purgehaplotigs using 10x data:

```
#!/bin/bash

#PBS -N 2020-05-02.PurgeHap.pbs
#PBS -V
#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
```

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

```

GENOME=./L_RNA_scaffolder.polished.fasta
BUSCO=../run_L_RNA_scaffolder.polished.busco/full_table_L_RNA_scaffolder.polished.busco.tsv
READS=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.3_StarlingNanopore/data/basecall/pass/filtered/filtered.fq
KMERREADS="/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/rawdata/HN00105164/HN00105164_10x_RawData_Outs/H2CYFCCX2/fastq_path/H2CYFCCX2/SV01/SV01_S1_L006_R*_001.fastq"
SCREENDB=/home/z5188231/programs/diploidocus-master/data/vecscreendb.fasta
PREFIX=L_RNA_scaffolder.polished.tidy

module add python/3.7.3 kat/2.4.2 perl/5.28.0 bedtools/2.27.1 R/3.5.3 samtools/1.10 purge_haplotigs/20190612 java/8u231-jre bbmap/38.51 minimap2/2.17 blast+/2.9.0 python/2.7.15

python /home/z3452659/slimsuitedev/tools/diploidocus.py -seqin $GENOME -runmode purgehap -basefile $PREFIX -busco $BUSCO -reads $READS kmerreads=\"$KMERREADS\" 10xtrim=T 10xtrim -forks 16 -screendb $SCREENDB pretrim=T
```

```
mkdir slimsute && cd $_
module load python/2.7.15
python /home/z3452659/slimsuitedev/tools/seqsuite.py summarise batchrun="../L_RNA_scaffolder.polished.tidy.purge.fasta" basefile=scaffolds dna newlog
```

```
#~~#    00:02:48      # ~~~~ Sequence Summary for L_RNA_scaffolder.polished.tidy.purge ~~~~ #
#SUM    00:03:28        Total number of sequences: 6,222
#SUM    00:03:28        Total length of sequences: 1,047,755,039
#SUM    00:03:28        Min. length of sequences: 917
#SUM    00:03:28        Max. length of sequences: 31,169,695
#SUM    00:03:28        Mean length of sequences: 168,395.22
#SUM    00:03:28        Median length of sequences: 2,199
#SUM    00:03:28        N50 length of sequences: 7,615,694
#SUM    00:03:28        L50 count of sequences: 37
#SUM    00:03:28        GC content: 41.73%
#SUM    00:03:28        Gap (N) length: 11,158,567 (1.06%)
```

Just quickly rerunning the above one so the HTML documentation can be generated with figures.

`dochtml=T/F`

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

```

GENOME=../L_RNA_scaffolder.polished.fasta
BUSCO=../run_L_RNA_scaffolder.polished.busco/full_table_L_RNA_scaffolder.polished.busco.tsv
READS=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.3_StarlingNanopore/data/basecall/pass/filtered/filtered.fq
KMERREADS="/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/rawdata/HN00105164/HN00105164_10x_RawData_Outs/H2CYFCCX2/fastq_path/H2CYFCCX2/SV01/SV01_S1_L006_R*_001.fastq"
SCREENDB=/home/z5188231/programs/diploidocus-master/data/vecscreendb.fasta
PREFIX=L_RNA_scaffolder.polished.tidy

module add python/3.7.3 kat/2.4.2 perl/5.28.0 bedtools/2.27.1 R/3.5.3 samtools/1.10 purge_haplotigs/20190612 java/8u231-jre bbmap/38.51 minimap2/2.17 blast+/2.9.0 python/2.7.15

python /home/z3452659/slimsuitedev/tools/diploidocus.py -seqin $GENOME -runmode purgehap -basefile $PREFIX -busco $BUSCO -reads $READS kmerreads=\"$KMERREADS\" 10xtrim=T 10xtrim -forks 16 -screendb $SCREENDB pretrim=T
```