Starling-May18
Projects/Katarina

PDF Version generated by

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

on

Jun 23, 2022 @03:56 PM NZST

Table of Contents

2020-03-28.ChromosomeAlignment

Katarina Stuart (z5188231@ad.unsw.edu.au) - Sep 11, 2020, 7:29 PM NZST

Aligning assembly to other chromosome level assemblies

Programs to try:

- RaGOO: https://github.com/malonge/RaGOO
- Ragout: https://fenderglass.github.io/Ragout/usage.html & https://github.com/fenderglass/Ragout
- PAFScaff (Slimsuite): https://github.com/slimsuite/pafscaff/blob/master/PAFScaff.md
- Satsuma: http://satsuma.sourceforge.net/
- Satsuma2: https://github.com/bioinfologics/satsuma2

Possible chromosome level reference passerine genomes (https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/):

1. Ficedula albicollis (collared flycatcher): https://www.ncbi.nlm.nih.gov/assembly/GCF_000247815.1
2. Taeniopygia guttata (zebra finch): https://www.ncbi.nlm.nih.gov/assembly/GCF_008822105.2/
3. Corvus moneduloides (New Caledonian crow)
4. Lonchura striata (white-rumped munia)
5. Passer domesticus (House sparrow)
6. Parus major (Great Tit)
7. Malurus cyaneus
8. Camarhynchus parvulus

RaGOO

RaGOO is a tool for coalescing genome assembly contigs into pseudochromosomes via minimap2 alignments to a closely related reference genome.

usage:
ragoo.py [-h] [-e <exclude.txt>] [-gff <annotations.gff>] [-m PATH] [-b] [-R <reads.fasta>] [-T sr] [-t 3] [-g 100] [-s] [-i 0.2] [-j <skip.txt>] [-C] <contigs.fasta> <reference.fasta>

order and orient contigs according to minimap2 alignments to a reference (v1.1)

positional arguments:
<contigs.fasta> fasta file with contigs to be ordered and oriented
<reference.fasta> reference fasta file

optional arguments:
-h, --help show this help message and exit
-e <exclude.txt> single column text file of reference headers to ignore
-gff <annotations.gff> lift-over gff features to chimera-broken contigs
-m PATH path to minimap2 executable
-b Break chimeric contigs
-R <reads.fasta> Turns on misassembly correction. Align provided reads to the contigs to aid misassembly correction. fastq or fasta allowed. Gzipped files allowed. Turns off '-b'.
-T sr Type of reads provided by '-R'. 'sr' and 'corr' accepted for short reads and error corrected long reads respectively.
-t 3 Number of threads when running minimap.
-g 100 Gap size for padding in pseudomolecules.
-s Call structural variants
-i 0.2 Minimum grouping confidence score needed to be localized.
-j <skip.txt> List of contigs to automatically put in chr0.
-C Write unplaced contigs individually instead of making a chr0

```
module load python/3.7.4
module load ragoo/1.11
module load minimap2/2.17
```

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/ragoo
```

```
ln -s /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/Diplodocus_tidy_Nala/purge_L_RNA_scaffolder.polished.tidy.1/L_RNA_scaffolder.polished.tidy.1.purge.fasta
ln -s /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/reference_assemblies/Taeniopygia_guttata/ncbi-genomes-2020-03-27/GCF_008822105.2_bTaeGut2.pat.W.v2_genomic.fna

ragoo.py -b -C -i 0.2 L_RNA_scaffolder.polished.tidy.1.purge.fasta GCF_008822105.2_bTaeGut2.pat.W.v2_genomic.fna
```

```
READS="/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/data/fastq/SV01_S1_L006_R*_001.fastq"
kmerreads=\"$READS\"

ragoo.py -R $READS -T sr -C -i 0.2 L_RNA_scaffolder.polished.tidy.1.purge.fasta GCF_008822105.2_bTaeGut2.pat.W.v2_genomic.fna
```

```
#the below started running. Above did not
ragoo.py -R \"$READS\" -T sr -C -i 0.2 L_RNA_scaffolder.polished.tidy.1.purge.fasta GCF_008822105.2_bTaeGut2.pat.W.v2_genomic.fna
```

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

```
ragoo.py -b L_RNA_scaffolder.polished.tidy.1.purge.fasta GCF_008822105.2_bTaeGut2.pat.W.v2_genomic.fna
#~~#    00:00:03        # ~~~~~~~~~~~~~~~ Sequence Summary for ragoo ~~~~~~~~~~~~~~~ #
#SUM    00:00:21        Total number of sequences: 120
#SUM    00:00:21        Total length of sequences: 1,048,367,551
#SUM    00:00:21        Min. length of sequences: 2,385
#SUM    00:00:21        Max. length of sequences: 150,341,846
#SUM    00:00:21        Mean length of sequences: 8,736,396.26
#SUM    00:00:21        Median length of sequences: 151,608
#SUM    00:00:21        N50 length of sequences: 71,537,503
#SUM    00:00:21        L50 count of sequences: 6
#SUM    00:00:21        GC content: 41.73%
#SUM    00:00:21        Gap (N) length: 11,778,567 (1.12%)
#SAVE   00:00:21        Table "summarise" saved to "scaffolds.summarise.tdt": 1 entries.

ragoo.py -b -C L_RNA_scaffolder.polished.tidy.1.purge.fasta GCF_008822105.2_bTaeGut2.pat.W.v2_genomic.fna
#~~#    00:00:03        # ~~~~~~~~~~~~~~~ Sequence Summary for ragoo ~~~~~~~~~~~~~~~ #
#SUM    00:00:21        Total number of sequences: 3,578
#SUM    00:00:21        Total length of sequences: 1,048,022,751
#SUM    00:00:21        Min. length of sequences: 32
#SUM    00:00:21        Max. length of sequences: 149,742,081
#SUM    00:00:21        Mean length of sequences: 292,907.42
#SUM    00:00:21        Median length of sequences: 1,381
#SUM    00:00:21        N50 length of sequences: 71,537,503
#SUM    00:00:21        L50 count of sequences: 6
#SUM    00:00:21        GC content: 41.73%
#SUM    00:00:21        Gap (N) length: 11,433,767 (1.09%)
#SAVE   00:00:21        Table "summarise" saved to "scaffolds.summarise.tdt": 1 entries.

ragoo.py -b -C -i 0.2 L_RNA_scaffolder.polished.tidy.1.purge.fasta GCF_008822105.2_bTaeGut2.pat.W.v2_genomic.fna
#~~#    00:00:03        # ~~~~~~~~~~~~~~~ Sequence Summary for ragoo ~~~~~~~~~~~~~~~ #
#SUM    00:00:21        Total number of sequences: 3,578
#SUM    00:00:21        Total length of sequences: 1,048,022,751
#SUM    00:00:21        Min. length of sequences: 32
#SUM    00:00:21        Max. length of sequences: 149,742,081
#SUM    00:00:21        Mean length of sequences: 292,907.42
#SUM    00:00:21        Median length of sequences: 1,381
#SUM    00:00:21        N50 length of sequences: 71,537,503
#SUM    00:00:21        L50 count of sequences: 6
#SUM    00:00:21        GC content: 41.73%
#SUM    00:00:21        Gap (N) length: 11,433,767 (1.09%)
#SAVE   00:00:21        Table "summarise" saved to "scaffolds.summarise.tdt": 1 entries.

ragoo.py -R \"$READS\" -T sr -C -i 0.2 L_RNA_scaffolder.polished.tidy.1.purge.fasta GCF_008822105.2_bTaeGut2.pat.W.v2_genomic.fna
#~~#    00:00:03        # ~~~~~~~~~~~~~~~ Sequence Summary for ragoo ~~~~~~~~~~~~~~~ #
#SUM    00:00:21        Total number of sequences: 3,573
#SUM    00:00:21        Total length of sequences: 1,048,012,351
#SUM    00:00:21        Min. length of sequences: 917
#SUM    00:00:21        Max. length of sequences: 150,340,946
#SUM    00:00:21        Mean length of sequences: 293,314.40
#SUM    00:00:21        Median length of sequences: 1,382
#SUM    00:00:21        N50 length of sequences: 71,537,277
#SUM    00:00:21        L50 count of sequences: 6
#SUM    00:00:21        GC content: 41.73%
#SUM    00:00:21        Gap (N) length: 11,423,367 (1.09%)
#SAVE   00:00:21        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 ../ragoo.fasta -o ragoo.busco -m genome -l ${BUSCOSET}/aves_odb9/ -c 32 -f
```

```
INFO    Results: pre scaffolding
INFO    C:94.3%[S:92.3%,D:2.0%],F:3.5%,M:2.2%,n:4915 for some reason this does down despite the fact that the numbers dont add up this way. Odd, but will proceed anyway.
INFO    4638 Complete BUSCOs (C)
INFO    4539 Complete and single-copy BUSCOs (S)
INFO    99 Complete and duplicated BUSCOs (D)
INFO    170 Fragmented BUSCOs (F)
```

INFO 107 Missing BUSCOs (M)
INFO 4915 Total BUSCO groups searched

INFO Results:
INFO C:94.2%[S:93.1%,D:1.1%],F:3.5%,M:2.3%,n:4915
INFO 4630 Complete BUSCOs (C)
INFO 4575 Complete and single-copy BUSCOs (S)
INFO 55 Complete and duplicated BUSCOs (D)
INFO 173 Fragmented BUSCOs (F)
INFO 112 Missing BUSCOs (M)
INFO 4915 Total BUSCO groups searched

Deleted as not the best version

RAGOUT:

Am stopping this as the setup is too confusing and not worth it. I call this intelligent and not lazy.

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/ragout
```

Installing:

```
conda install -c bioconda ragout
module load cmake/3.14.5 gcc/7.3.0

git clone https://github.com/fenderglass/Ragout.git
cd Ragout
python setup.py build
pip install -r requirements.txt --user
python scripts/install-sibelia.py

module load hal/20190129
module load python/3.6.5

!Might need: python-networkx == 2.2 & GNU make
```

Cactus install (https://github.com/ComparativeGenomicsToolkit/cactus):

```
python3 -m pip install virtualenv
```

```
virtualenv -p python3.6 cactus_env
```

```
source cactus_env/bin/activate
```

You can always exit out of the virtualenv by running `deactivate`

```
pip install --upgrade toil[all]
```

Configuration (Recipe) File:

```
.references = TGut
.target = SVul
```

```
TGut.fasta = /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/reference_assemblies/Taeniopygia_guttata/ncbi-genomes-2020-03-27/GCF_008822105.2_bTaeGut2.pat.W.v2_genomic.fna
SVul.fasta =  /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/Diplodocus_tidy_Nala/purge_L_RNA_scaffolder.polished.tidy.1/L_RNA_scaffolder.polished.tidy.1.purge.fasta
```

Satsuma2:

Map your scaffolds or contigs onto chromosome coordinates via synteny! To do so, run

./Chromosemble -t <reference> -q <your_scaffolds> -o <output_dir>

The full list of options is:

-t<string> : target fasta file (in chromosome coordinates)
-q<string> : query fasta file (the assembly)
-o<string> : output directory
-n<int> : number of CPUs (for full Satsuma run) (def=25)
-thorough<bool> : runs a full Satsuma alignment (slow!!) (def=0)

-pseudochr<bool> : maps scaffolds into chromosomes (def=0)
-s<bool> : run SatsumaSynteny at the end (def=0)

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/
```

Setting up program:

```
git clone https://github.com/bioinfologics/satsuma2
module load cmake/3.14.5 gcc/7.3.0
CMake CMakeLists.txt
```

```
chmod u+r+x *
cmake CMakeCache.txt
cmake_install.cmake
make -f Makefile
```

```
export PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2:$PATH
export PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2/bin:$PATH
export SATSUMA2_PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2/bin
```

Running Satsuma2:

```
Chromosemble -t GCF_008822105.2_bTaeGut2.pat.W.v2_genomic.fna -q L_RNA_scaffolder.polished.tidy.1.purge.fasta -o Chromosemble.fasta
```

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

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/Chromosemble.fasta/slimsute
```

```
#~~#    00:02:41       # ~~~~~~~~ Sequence Summary for pseudochromosomes ~~~~~~~~ #
#SUM    00:03:23       Total number of sequences: 1,628
#SUM    00:03:23       Total length of sequences: 1,049,829,390
#SUM    00:03:23       Min. length of sequences: 927
#SUM    00:03:23       Max. length of sequences: 151,927,750
#SUM    00:03:23       Mean length of sequences: 644,858.35
#SUM    00:03:23       Median length of sequences: 1,337
#SUM    00:03:23       N50 length of sequences: 72,525,610
#SUM    00:03:23       L50 count of sequences: 5
#SUM    00:03:23       GC content: 41.73%
#SUM    00:03:23       Gap (N) length: 13,240,406 (1.26%)
#LOAD   00:03:23       Load sequences from ../superscaffolds.fasta
#SEQ    00:06:05       4,886 of 4,886 sequences loaded from ../superscaffolds.fasta (Format: fas).
#INDEX  00:06:05       Index file ../superscaffolds.fasta.index made
#FILT   00:06:05       4,886 of 4,886 sequences retained.
#~~#    00:06:05       # ~~~~~~~~~ Sequence Summary for superscaffolds ~~~~~~~~~~ #
#SUM    00:06:46       Total number of sequences: 4,886
#SUM    00:06:46       Total length of sequences: 1,047,881,051
#SUM    00:06:46       Min. length of sequences: 917
#SUM    00:06:46       Max. length of sequences: 52,382,189
#SUM    00:06:46       Mean length of sequences: 214,466.04
#SUM    00:06:46       Median length of sequences: 1,712
#SUM    00:06:46       N50 length of sequences: 14,504,343
#SUM    00:06:46       L50 count of sequences: 22
#SUM    00:06:46       GC content: 41.73%
#SUM    00:06:46       Gap (N) length: 11,292,067 (1.08%)
```

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

```
INFO    Results:
INFO    C:94.6%[S:93.5%,D:1.1%],F:3.1%,M:2.3%,n:4915
INFO    4649 Complete BUSCOs (C)
INFO    4595 Complete and single-copy BUSCOs (S)
INFO    54 Complete and duplicated BUSCOs (D)
INFO    154 Fragmented BUSCOs (F)
INFO    112 Missing BUSCOs (M)
INFO    4915 Total BUSCO groups searched
```

Visualisation

- BlockDisplaySatsuma: takes a satsuma summary file and writes displayable blocks in MizBee format, see http://www.cs.utah.edu/~miriah/mizbee/Overview.html for how to display this using the MizBee Synteny Browser.
- ChromosomePaint: generates a comparative chromosome view in postscript format from the MizBee file generated by BlockDisplaySatsuma.

```
SatsumaSynteny2 -t GCF_008822105.2_bTaeGut2.pat.W.v2_genomic.fna -q L_RNA_scaffolder.polished.tidy.1.purge.fasta -o Synteny2.summary
```

RERUN this (48 queue submit or interactive) when I have the final version that I want to align.

Available arguments:

-i<string> : MizBee file
-o<string> : outfile (post-script)
-d<double> : dot size (def=1)
-s<double> : scale (def=60000)
-t<int> : target id (def=-1)
-d<bool> : print indivisual matchs (def=0)
-f<bool> : forward only (def=0)

bin/ChromosomePaint

Satsuma 2 on the final dipcycle genome

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/
```

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

```
export PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2:$PATH
export PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2/bin:$PATH
export SATSUMA2_PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2/bin
```

```
Chromosemble -t GCF_008822105.2_bTaeGut2.pat.W.v2_genomic.fna -q L_RNA_scaffolder.polished.tidy.diploidocus.fasta -o Chromosemble.diploidocus.fasta
```

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

```
#~~#    00:01:52        # ~~~~~~~~ Sequence Summary for pseudochromosomes ~~~~~~~~ #
#SUM    00:02:24        Total number of sequences: 334
#SUM    00:02:24        Total length of sequences: 1,035,260,756
#SUM    00:02:24        Min. length of sequences: 1,001
#SUM    00:02:24        Max. length of sequences: 161,466,563
#SUM    00:02:24        Mean length of sequences: 3,099,583.10
#SUM    00:02:24        Median length of sequences: 2,527
#SUM    00:02:24        N50 length of sequences: 72,051,062
#SUM    00:02:24        L50 count of sequences: 5
#SUM    00:02:24        GC content: 41.58%
#SUM    00:02:24        Gap (N) length: 10,957,698 (1.06%)
#LOAD   00:02:24        Load sequences from ../superscaffolds.fasta
#SEQ    00:04:15        1,680 of 1,680 sequences loaded from ../superscaffolds.fasta (Format: fas).
#INDEX  00:04:15        Index file ../superscaffolds.fasta.index made
#FILT   00:04:15        1,680 of 1,680 sequences retained.
#~~#    00:04:15        # ~~~~~~~~~~ Sequence Summary for superscaffolds ~~~~~~~~~~ #
#SUM    00:04:46        Total number of sequences: 1,680
#SUM    00:04:46        Total length of sequences: 1,034,487,168
#SUM    00:04:46        Min. length of sequences: 1,001
#SUM    00:04:46        Max. length of sequences: 91,735,643
#SUM    00:04:46        Mean length of sequences: 615,766.17
#SUM    00:04:46        Median length of sequences: 12,715
#SUM    00:04:46        N50 length of sequences: 16,532,297
#SUM    00:04:46        L50 count of sequences: 17
#SUM    00:04:46        GC content: 41.58%
#SUM    00:04:46        Gap (N) length: 10,184,110 (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 ../pseudochromosomes.fasta -o pseudochromosomes.busco -m genome -l ${BUSCOSET}/aves_odb9/ -c 32 -f
```

```
INFO    Results:
INFO    C:94.2%[S:93.2%,D:1.0%],F:3.3%,M:2.5%,n:4915
INFO    4629 Complete BUSCOs (C)
INFO    4579 Complete and single-copy BUSCOs (S)
INFO    50 Complete and duplicated BUSCOs (D)
INFO    164 Fragmented BUSCOs (F)
INFO    122 Missing BUSCOs (M)
INFO    4915 Total BUSCO groups searched
```

Deleted as not the best version

Satsuma 2 on the svulgaris-10x-550M-sub80.pri

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/
```

```
GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/assembly/Diploidocus/svulgaris-10x-550M-sub80.pri.fasta
```

```
export PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2:$PATH
export PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2/bin:$PATH
export SATSUMA2_PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2/bin
```

```
Chromosemble -t GCF_008822105.2_bTaeGut2.pat.W.v2_genomic.fna -q $GENOME -o Chromosemble.svulgaris-10x-550M-sub80.pri.fasta
```

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

```
#~~#    00:02:37        # ~~~~~~~~ Sequence Summary for pseudochromosomes ~~~~~~~~ #
#SUM    00:03:20        Total number of sequences: 4,157
#SUM    00:03:20        Total length of sequences: 1,044,908,401
#SUM    00:03:20        Min. length of sequences: 1,000
#SUM    00:03:20        Max. length of sequences: 150,712,711
#SUM    00:03:20        Mean length of sequences: 251,361.17
#SUM    00:03:20        Median length of sequences: 1,367
#SUM    00:03:20        N50 length of sequences: 73,194,681
#SUM    00:03:20        L50 count of sequences: 5
#SUM    00:03:20        GC content: 41.64%
#SUM    00:03:20        Gap (N) length: 11,637,699 (1.11%)
#LOAD   00:03:20        Load sequences from ../superscaffolds.fasta
#SEQ    00:05:54        13,845 of 13,845 sequences loaded from ../superscaffolds.fasta (Format: fas).
#INDEX  00:05:54        Index file ../superscaffolds.fasta.index made
#FILT   00:05:54        13,845 of 13,845 sequences retained.
#~~#    00:05:54        # ~~~~~~~~~~ Sequence Summary for superscaffolds ~~~~~~~~~~ #
#SUM    00:06:36        Total number of sequences: 13,845
#SUM    00:06:36        Total length of sequences: 1,040,565,892
#SUM    00:06:36        Min. length of sequences: 1,000
#SUM    00:06:36        Max. length of sequences: 39,562,901
#SUM    00:06:36        Mean length of sequences: 75,158.24
#SUM    00:06:36        Median length of sequences: 1,716
#SUM    00:06:36        N50 length of sequences: 5,384,266
#SUM    00:06:36        L50 count of sequences: 54
#SUM    00:06:36        GC content: 41.64%
#SUM    00:06:36        Gap (N) length: 7,295,190 (0.70%)
#SAVE   00:06:36        Table "summarise" saved to "scaffolds.summarise.tdt": 2 entries.
```

Deleted as not the best version

Satsuma 2 on the scaffolds_gapfilled_FINAL.fasta

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/
```

```
GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/SSPACE_GapfinisherV2/scaffolds_gapfilled_FINAL.fasta
```

```
export PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2:$PATH
export PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2/bin:$PATH
export SATSUMA2_PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2/bin
```

```
Chromosemble -t GCF_008822105.2_bTaeGut2.pat.W.v2_genomic.fna -q $GENOME -o Chromosemble.scaffolds_gapfilled_FINAL.fasta
```

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

```
#~~#    00:02:48       # ~~~~~~~~ Sequence Summary for pseudochromosomes ~~~~~~~~ #
#SUM    00:03:32       Total number of sequences: 1,978
#SUM    00:03:32       Total length of sequences: 1,069,752,029
#SUM    00:03:32       Min. length of sequences: 977
#SUM    00:03:32       Max. length of sequences: 153,044,858
#SUM    00:03:32       Mean length of sequences: 540,825.09
#SUM    00:03:32       Median length of sequences: 1,348
#SUM    00:03:32       N50 length of sequences: 72,902,629
#SUM    00:03:32       L50 count of sequences: 5
#SUM    00:03:32       GC content: 41.82%
#SUM    00:03:32       Gap (N) length: 14,029,129 (1.31%)
#LOAD   00:03:32        Load sequences from ../superscaffolds.fasta
#SEQ    00:06:25       6,225 of 6,225 sequences loaded from ../superscaffolds.fasta (Format: fas).
#INDEX  00:06:26        Index file ../superscaffolds.fasta.index made
#FILT   00:06:26       6,225 of 6,225 sequences retained.
#~~#    00:06:26       # ~~~~~~~~~~ Sequence Summary for superscaffolds ~~~~~~~~~~ #
#SUM    00:07:09       Total number of sequences: 6,225
#SUM    00:07:09       Total length of sequences: 1,067,476,876
#SUM    00:07:09       Min. length of sequences: 977
#SUM    00:07:09       Max. length of sequences: 49,791,553
#SUM    00:07:09       Mean length of sequences: 171,482.23
#SUM    00:07:09       Median length of sequences: 1,844
#SUM    00:07:09       N50 length of sequences: 12,295,759
#SUM    00:07:09       L50 count of sequences: 24
#SUM    00:07:09       GC content: 41.82%
#SUM    00:07:09       Gap (N) length: 11,753,976 (1.10%)
```

Deleted as not the best version

Satsuma 2 on the clustered_L_RNA_scaffolder.polished.hq.fasta

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/
```

```
GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/Pilon/bwa-aligned_scaffolded/L_RNA_scaffolder.polished.fasta
```

```
export PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2:$PATH
export PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2/bin:$PATH
export SATSUMA2_PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2/bin
```

```
Chromosemble -t GCF_008822105.2_bTaeGut2.pat.W.v2_genomic.fna -q $GENOME -o Chromosemble.L_RNA_scaffolder.polished.fasta
```

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

```
#~~#    00:02:12       # ~~~~~~~~ Sequence Summary for pseudochromosomes ~~~~~~~~ #
#SUM    00:02:53       Total number of sequences: 1,976
#SUM    00:02:53       Total length of sequences: 1,069,498,029
#SUM    00:02:53       Min. length of sequences: 927
#SUM    00:02:53       Max. length of sequences: 152,803,141
#SUM    00:02:53       Mean length of sequences: 541,243.94
#SUM    00:02:53       Median length of sequences: 1,348
#SUM    00:02:53       N50 length of sequences: 72,895,019
#SUM    00:02:53       L50 count of sequences: 5
#SUM    00:02:53       GC content: 41.82%
#SUM    00:02:53       Gap (N) length: 13,822,972 (1.29%)
#LOAD   00:02:53        Load sequences from ../superscaffolds.fasta
#SEQ    00:05:06       6,163 of 6,163 sequences loaded from ../superscaffolds.fasta (Format: fas).
#INDEX  00:05:06        Index file ../superscaffolds.fasta.index made
#FILT   00:05:06       6,163 of 6,163 sequences retained.
#~~#    00:05:06       # ~~~~~~~~~~ Sequence Summary for superscaffolds ~~~~~~~~~~ #
#SUM    00:05:46       Total number of sequences: 6,163
#SUM    00:05:46       Total length of sequences: 1,067,232,500
#SUM    00:05:46       Min. length of sequences: 842
#SUM    00:05:46       Max. length of sequences: 74,348,001
#SUM    00:05:46       Mean length of sequences: 173,167.69
```

```
#SUM    00:05:46        Median length of sequences: 1,831
#SUM    00:05:46        N50 length of sequences: 14,505,999
#SUM    00:05:46        L50 count of sequences: 21
#SUM    00:05:46        GC content: 41.82%
#SUM    00:05:46        Gap (N) length: 11,557,443 (1.08%)
```

```
#!/bin/bash

#PBS -N 2020-05-01.BUSCO.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/chromosome_alignment/satsuma2/Chromosemble.L_RNA_scaffolder.polished.fasta/slimsute

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

```
C:94.5%[S:93.3%,D:1.2%],F:3.3%,M:2.2%,n:4915

    4642    Complete BUSCOs (C)
    4584    Complete and single-copy BUSCOs (S)
    58      Complete and duplicated BUSCOs (D)
    162     Fragmented BUSCOs (F)
    111     Missing BUSCOs (M)
    4915    Total BUSCO groups searched
```

Deleted as not the best version

FINAL VERSION: L_RNA_scaffolder.polished.tidy.purge

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/
```

```
GENOME=/srv/scratch/z5188231/KStuart.Starling-
Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/Diplodocus_tidy_all/Purgehap/purge_L_RNA_scaffolder.polished.tidy/L_RNA_scaffolder.polished.tidy.purge.fasta
```

```
export PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2:$PATH
export PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2/bin:$PATH
export SATSUMA2_PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2/bin
```

```
Chromosemble -t GCF_008822105.2_bTaeGut2.pat.W.v2_genomic.fna -q $GENOME -o Chromosemble.L_RNA_scaffolder.polished.tidy.purge.fasta
```

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

```
#~~#    00:02:44        # ~~~~~~~~ Sequence Summary for pseudochromosomes ~~~~~~~~ #
#SUM    00:03:26        Total number of sequences: 1,628
#SUM    00:03:26        Total length of sequences: 1,049,838,585
#SUM    00:03:26        Min. length of sequences: 927
#SUM    00:03:26        Max. length of sequences: 151,927,750
#SUM    00:03:26        Mean length of sequences: 644,864.00
#SUM    00:03:26        Median length of sequences: 1,337
#SUM    00:03:26        N50 length of sequences: 72,525,610
#SUM    00:03:26        L50 count of sequences: 5
#SUM    00:03:26        GC content: 41.73%
#SUM    00:03:26        Gap (N) length: 13,242,113 (1.26%)
#SEQ    00:06:08        4,887 of 4,887 sequences loaded from ../superscaffolds.fasta (Format: fas).
#INDEX  00:06:08        Index file ../superscaffolds.fasta.index made
#FILT   00:06:08        4,887 of 4,887 sequences retained.
#~~#    00:06:08        # ~~~~~~~~~ Sequence Summary for superscaffolds ~~~~~~~~~~ #
```

```
#SUM    00:06:48    Total number of sequences: 4,887
#SUM    00:06:48    Total length of sequences: 1,047,888,539
#SUM    00:06:48    Min. length of sequences: 917
#SUM    00:06:48    Max. length of sequences: 52,382,189
#SUM    00:06:48    Mean length of sequences: 214,423.68
#SUM    00:06:48    Median length of sequences: 1,712
#SUM    00:06:48    N50 length of sequences: 14,504,343
#SUM    00:06:48    L50 count of sequences: 22
#SUM    00:06:48    GC content: 41.73%
#SUM    00:06:48    Gap (N) length: 11,292,067 (1.08%)
```

```
#!/bin/bash

#PBS -N 2020-05-04.BUSCO.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/chromosome_alignment/satsuma2/Chromosemble.L_RNA_scaffolder.polished.tidy.purge.fasta/slimsute

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

```
# BUSCO was run in mode: genome

    C:94.6%[S:93.5%,D:1.1%],F:3.1%,M:2.3%,n:4915

    4649    Complete BUSCOs (C)
    4595    Complete and single-copy BUSCOs (S)
    54      Complete and duplicated BUSCOs (D)
    154     Fragmented BUSCOs (F)
    112     Missing BUSCOs (M)
    4915    Total BUSCO groups searched
```

Rename and simplify fasta titles for future analysis

```
cp pseudochromosomes.fasta Sturnus_vulgaris_2.3.fasta
```

```
perl /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/data_2020/adv_repeats/programs/simplifyFastaHeaders.pl Sturnus_vulgaris_2.3.fasta starling1 Sturnus_vulgaris_2.3.simp.fasta Sturnus_vulgaris_2.3.map
```

```
perl /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.4_GenomeAnnotation/data_2020/adv_repeats/programs/simplifyFastaHeaders.pl Sturnus_vulgaris_2.3.fasta starling Sturnus_vulgaris_2.3.1.simp.fasta Sturnus_vulgaris_2.3.1.map
```

For linking to directories:

```
ln -s /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/Chromosemble.L_RNA_scaffolder.polished.tidy.purge.fasta/Sturnus_vulgaris_2.3.simp.fasta .
```

Visualisation:

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/
```

```
GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/nanopore.scaffolding/Diplodocus_tidy_all/Purgehap/purge_L_RNA_scaffolder.polished.tidy/L_RNA_scaffolder.polished.tidy.purge.fasta
```

```
module load cmake/3.14.5 gcc/7.3.0
```

```
export PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2:$PATH
export PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2/bin:$PATH
export SATSUMA2_PATH=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv3_Genome/Sv3.2_Starling10x/chromosome_alignment/satsuma2/satsuma2/bin
```

```
SatsumaSynteny2 -t GCF_008822105.2_bTaeGut2.pat.W.v2_genomic.fna -q $GENOME -o Synteny2.L_RNA_scaffolder.polished.tidy.purge.summary -threads 6
```

TIME SPENT WORKING: 148753
Joining Workqueue thread
SATSUMA: all done, date and time: 2020/07/01 06:15:28

```
BlockDisplaySatsuma -i Synteny2.L_RNA_scaffolder.polished.tidy.purge.summary -t GCF_008822105.2_bTaeGut2.pat.W.v2_genomic.fna -q $GENOME
> BlockDisplaySatsuma.out
```

```
ChromosomePaint -i BlockDisplaySatsuma.out -o ChromosomePaint.ps
```

-i<string> : MizBee file
-o<string> : outfile (post-script)
-d<double> : dot size (def=1)
-s<double> : scale (def=60000)
-t<int> : target id (def=-1)
-d<bool> : print indivisual matchs (def=0)
-f<bool> : forward only (def=0)