Starling-May18
Projects/Katarina Stuart/KStuart.Starling-Aug18/At1_Genome/Assembly/2022-11-25.Curation

PDF Version generated by

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

on

Aug 23, 2023 @09:32 AM NZST

Table of Contents

2022-11-25.Curation

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

Myna Genome Curation

Genome versions (starling)

```
GENOME_VERSION1=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_sept/genome_QC/Assembly_ONT_noalt_scaf_medaka.fasta
GENOME_VERSION2=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_sept/genome_QC/Assembly_PacBio_ONT_medaka.fasta
```

Chose to proceed with Assembly_ONT_noalt_scaf_medaka.fasta

SOME PREAMBLE:

Busco

```
#!/bin/bash -e

#SBATCH --job-name=2022_11_25.genome_busco_ONT.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-24:00:00
#SBATCH --mem=50GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=8
#SBATCH --profile task

GENOME_VERSION1=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_sept/genome_QC/Assembly_ONT_noalt_scaf_medaka.fasta
DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/resources/genomes

module load BBMap/38.81-gimkl-2020a

#remove contig tail, with min length 1,500 bp
reformat.sh in=$GENOME_VERSION1 out=${DIR}/$(basename $GENOME_VERSION1 .fasta)_trimmed.fasta minlength=1500

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/genome_stats_summary/busco

module purge
module load BUSCO/5.3.2-gimkl-2020a
busco  -i ${DIR}/$(basename $GENOME_VERSION1 .fasta)_trimmed.fasta -o $(basename $GENOME_VERSION1 .fasta)_trimmed -m genome -l aves_odb10  -c 8 -f
```

POST TRIM:

C:96.6%[S:96.2%,D:0.4%],F:0.9%,M:2.5%,n:8338

8062 Complete BUSCOs (C)

8025 Complete and single-copy BUSCOs (S)

37 Complete and duplicated BUSCOs (D)

79 Fragmented BUSCOs (F)

197 Missing BUSCOs (M)

8338 Total BUSCO groups searched

PRE TRIM (?): /nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_sept/Basecalling_guppy6.2.1/busco_myna_ontraw_noalts_scafs_raw_241022

C:96.2%[S:95.8%,D:0.4%],F:1.1%,M:2.7%,n:8338

8024 Complete BUSCOs (C)

7987 Complete and single-copy BUSCOs (S)

37 Complete and duplicated BUSCOs (D)

93 Fragmented BUSCOs (F)

221 Missing BUSCOs (M)

8338 Total BUSCO groups searched

Seqsuite

```
#!/bin/bash -e

#SBATCH --job-name=2022_11_28.genome_seqsuite_ONT.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=5GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

module purge
module load Python/2.7.14-gimkl-2017a

GENOME_VERSION1=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_sept/genome_QC/Assembly_ONT_noalt_scaf_medaka.fasta
DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/myna_genome_versions

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/genome_stats_summary/seqsuite

python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/tools/seqsuite.py -seqin $GENOME_VERSION1 -summarise -dna

python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/tools/seqsuite.py -seqin ${DIR}/$(basename $GENOME_VERSION1 .fasta)_trimmed.fasta -summarise -dna
```

```
#~~#    00:00:00        # ~~~~ Sequence Summary for Assembly_ONT_noalt_scaf_medaka ~~~~ #
#SUM    00:00:36        Total number of sequences: 1,648
#SUM    00:00:36        Total length of sequences: 1,046,266,241
#SUM    00:00:36        Min. length of sequences: 13
#SUM    00:00:36        Max. length of sequences: 35,054,366
#SUM    00:00:36        Mean length of sequences: 634,870.29
#SUM    00:00:36        Median length of sequences: 2,253
#SUM    00:00:36        N50 length of sequences: 11,267,561
#SUM    00:00:36        L50 count of sequences: 29
#SUM    00:00:36        Total number of contigs: 1,648
#SUM    00:00:36        GC content: 41.87%
#SUM    00:00:36        N bases: 0 (0.00%)
#SUM    00:00:36        Gap (10+ N) length: 0 (0.00%)
#SUM    00:00:36        Gap (10+ N) count: 0

#~~#    00:02:21        # ~~~~ Sequence Summary for Assembly_ONT_noalt_scaf_medaka_trimmed ~~~~ #
#SUM    00:03:38        Total number of sequences: 991
#SUM    00:03:38        Total length of sequences: 1,045,720,465
#SUM    00:03:38        Min. length of sequences: 1,508
#SUM    00:03:38        Max. length of sequences: 35,054,366
#SUM    00:03:38        Mean length of sequences: 1,055,217.42
#SUM    00:03:38        Median length of sequences: 6,022
#SUM    00:03:38        N50 length of sequences: 11,267,561
#SUM    00:03:38        L50 count of sequences: 29
#SUM    00:03:38        Total number of contigs: 991
#SUM    00:03:38        GC content: 41.86%
#SUM    00:03:38        N bases: 0 (0.00%)
#SUM    00:03:38        Gap (10+ N) length: 0 (0.00%)
#SUM    00:03:38        Gap (10+ N) count: 0
```

Reordering contigs for IGV browsing

using the association table from the dgenies alignment to zebra finch genome

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/myna_genome_versions/igv_contig_reordering

#create contig order
DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/dgenies
cut -f1 ${DIR}/Assembly_ONT_noalt_scaf_medaka_GCF_003957565.2_bTaeGut1.4.pri_genomic.fna.fasta_assoc.tsv | tail -n +2 > Tgut_ONT_noalt_scaf_contigorder.txt

#reorder assembly
module load SAMtools/1.15.1-GCC-11.3.0
GENOME_VERSION1=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_sept/genome_QC/Assembly_ONT_noalt_scaf_medaka.fasta
#samtools faidx ${GENOME_VERSION1} #genome previously indexed
samtools faidx ${GENOME_VERSION1} $(cat Tgut_ONT_noalt_scaf_contigorder.txt) > Assembly_ONT_noalt_scaf_medaka.reordered.fasta
```

ONT & PACBIO:

Busco

```
#!/bin/bash -e

#SBATCH --job-name=2022_11_25.genome_busco_PB.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-24:00:00
#SBATCH --mem=50GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=8
#SBATCH --profile task

GENOME_VERSION2=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_sept/genome_QC/Assembly_PacBio_ONT_medaka.fasta
DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/resources/genomes

module load BBMap/38.81-gimkl-2020a

#remove contig tail, with min length 1,500 bp
reformat.sh in=$GENOME_VERSION2 out=${DIR}/$(basename $GENOME_VERSION2 .fasta)_trimmed.fasta minlength=1500

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/genome_stats_summary/busco

module purge
module load BUSCO/5.3.2-gimkl-2020a
busco  -i ${DIR}/$(basename $GENOME_VERSION2 .fasta)_trimmed.fasta -o $(basename $GENOME_VERSION2 .fasta)_trimmed -m genome -l aves_odb10  -c 8 -f
```

Seqsuite

```
#!/bin/bash -e

#SBATCH --job-name=2022_11_28.genome_seqsuite_PB.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=5GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

module purge
module load Python/2.7.14-gimkl-2017a

GENOME_VERSION2=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_sept/genome_QC/Assembly_PacBio_ONT_medaka.fasta
DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/myna_genome_versions

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/genome_stats_summary/seqsuite

python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/tools/seqsuite.py -seqin $GENOME_VERSION2 -summarise -dna
```

```
python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/tools/seqsuite.py -seqin ${DIR}/$(basename $GENOME_VERSION2 .fasta)_trimmed.fasta -summarise -dna
```

```
#~~#    00:00:00        # ~~~~ Sequence Summary for Assembly_PacBio_ONT_medaka ~~~~ #
#SUM    00:00:39        Total number of sequences: 6,119
#SUM    00:00:39        Total length of sequences: 1,091,022,534
#SUM    00:00:39        Min. length of sequences: 9
#SUM    00:00:39        Max. length of sequences: 47,758,989
#SUM    00:00:39        Mean length of sequences: 178,300.79
#SUM    00:00:39        Median length of sequences: 3,278
#SUM    00:00:39        N50 length of sequences: 10,108,235
#SUM    00:00:39        L50 count of sequences: 32
#SUM    00:00:39        Total number of contigs: 6,119
#SUM    00:00:39        GC content: 42.04%
#SUM    00:00:39        N bases: 0 (0.00%)
#SUM    00:00:39        Gap (10+ N) length: 0 (0.00%)
#SUM    00:00:39        Gap (10+ N) count: 0

#~~#    00:02:18        # ~~~~ Sequence Summary for Assembly_PacBio_ONT_medaka_trimmed ~~~~ #
#SUM    00:03:39        Total number of sequences: 3,995
#SUM    00:03:39        Total length of sequences: 1,089,310,912
#SUM    00:03:39        Min. length of sequences: 1,503
#SUM    00:03:39        Max. length of sequences: 47,758,989
#SUM    00:03:39        Mean length of sequences: 272,668.56
#SUM    00:03:39        Median length of sequences: 7,761
#SUM    00:03:39        N50 length of sequences: 10,108,235
#SUM    00:03:39        L50 count of sequences: 32
#SUM    00:03:39        Total number of contigs: 3,995
#SUM    00:03:39        GC content: 42.03%
#SUM    00:03:39        N bases: 0 (0.00%)
#SUM    00:03:39        Gap (10+ N) length: 0 (0.00%)
#SUM    00:03:39        Gap (10+ N) count: 0
```

Reordering contigs for IGV browsing

using the association table from the dgenies alignment to zebra finch genome

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/myna_genome_versions/igv_contig_reordering

#create contig order
DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/dgenies
cut -f1 ${DIR}/Assembly_PacBio_ONT_medaka_GCF_003957565.2_bTaeGut1.4.pri_genomic.fna.fasta_assoc.tsv | tail -n +2 > Tgut_PacBio_ONT_contigorder.txt

#reorder assembly
module load SAMtools/1.15.1-GCC-11.3.0
GENOME_VERSION2=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_sept/genome_QC/Assembly_PacBio_ONT_medaka.fasta
#samtools faidx ${GENOME_VERSION2} #genome previously indexed
samtools faidx ${GENOME_VERSION2} $(cat Tgut_PacBio_ONT_contigorder.txt) > Assembly_PacBio_ONT_medaka.reordered.fasta
```

RagTag:

For breaking scaffolds

https://github.com/malonge/RagTag/wiki/correct

installation

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs
echo "export PATH=\$PATH:\/nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/bin" >> $HOME/.bashrc # add to .bashrc
source $HOME/.bashrc
conda init
conda install -c bioconda ragtag
```

```
conda create -n ragtag tagtag#didn't work
#ragtag .py's here for some reason? Not sure why they installed as single executables and not as an env.
/nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/bin
```

Working with the ONT_noalt_scaff assembly

```
#!/bin/bash -e

#SBATCH --job-name=2022_12_05.RagTag.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-48:00:00
#SBATCH --mem=12GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=8
#SBATCH --profile task

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/ragtag

REF_TGUT=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/GCF_003957565.2_bTaeGut1.4.pri/GCF_003957565.2_bTaeGut1.4.pri_genomic.fna
REF_SVUL=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/Sturnus_vulgaris_2.3.1.simp.fasta
GENOME=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_sept/genome_QC/Assembly_ONT_noalt_scaf_medaka.fasta

#To the zebra finch genome
#validation reads:
#/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_sept/Basecalling_guppy6.2.1/pilon_polishing/Myna_10x_processed_R1_val_1.fq.gz
#/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_sept/Basecalling_guppy6.2.1/pilon_polishing/Myna_10x_processed_R2_val_2.fq.gz
ragtag.py correct ${REF_TGUT} ${GENOME} -t 8 -o ./ragtag_output_tgut_noval
ragtag.py correct ${REF_SVUL} ${GENOME} -t 8 -o ./ragtag_output_svul_noval
ragtag.py correct ${REF_TGUT} ${GENOME} -t 8 -F validation_reads.txt -T sr -o ./ragtag_output_tgut

#To the zebra finch genome
#ragtag.py correct ${REF_SVUL} ${GENOME} -t 8 -o ./ragtag_output_svul
```

Seqsuite of the new assemblies

```
#!/bin/bash -e

#SBATCH --job-name=2022_12_13.genome_seqsuite_ragtag.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=5GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

module purge
module load Python/2.7.14-gimkl-2017a

GENOME_VERSION1=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/ragtag/ragtag_output_tgut/ragtag.correct.fasta

DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/myna_genome_versions

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/genome_stats_summary/seqsuite

python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/tools/seqsuite.py -seqin $GENOME_VERSION1 -summarise -dna

python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/tools/seqsuite.py -seqin ${DIR}/$(basename $GENOME_VERSION1 .fasta)_trimmed.fasta -summarise -dna
```

```
#~~#  00:00:00  # ~~~~ Sequence Summary for Assembly_ONT_noalt_scaf_medaka ~~~~ #
#SUM  00:00:36  Total number of sequences: 1,648
#SUM  00:00:36  Total length of sequences: 1,046,266,241
#SUM  00:00:36  Min. length of sequences: 13
#SUM  00:00:36  Max. length of sequences: 35,054,366
#SUM  00:00:36  Mean length of sequences: 634,870.29
```

```
#SUM    00:00:36        Median length of sequences: 2,253
#SUM    00:00:36        N50 length of sequences: 11,267,561
#SUM    00:00:36        L50 count of sequences: 29
#SUM    00:00:36        Total number of contigs: 1,648
#SUM    00:00:36        GC content: 41.87%
#SUM    00:00:36        N bases: 0 (0.00%)
#SUM    00:00:36        Gap (10+ N) length: 0 (0.00%)
#SUM    00:00:36        Gap (10+ N) count: 0

#~~#    00:02:21        # ~~~~ Sequence Summary for Assembly_ONT_noalt_scaf_medaka_trimmed ~~~~ #
#SUM    00:03:38        Total number of sequences: 991
#SUM    00:03:38        Total length of sequences: 1,045,720,465
#SUM    00:03:38        Min. length of sequences: 1,508
#SUM    00:03:38        Max. length of sequences: 35,054,366
#SUM    00:03:38        Mean length of sequences: 1,055,217.42
#SUM    00:03:38        Median length of sequences: 6,022
#SUM    00:03:38        N50 length of sequences: 11,267,561
#SUM    00:03:38        L50 count of sequences: 29
#SUM    00:03:38        Total number of contigs: 991
#SUM    00:03:38        GC content: 41.86%
#SUM    00:03:38        N bases: 0 (0.00%)
#SUM    00:03:38        Gap (10+ N) length: 0 (0.00%)
#SUM    00:03:38        Gap (10+ N) count: 0
```

FINAL CURATION:

Curation process:

1) Manual breaking round 1: use mapped tracks to break most obvious misassemblies. Ragtag informed as well (from the validates ones)
2) realign the plots and visualise in dgenies
3) another road of breaking, using the previous mapped tracks
4) Repeat visualisation and breaking as needed
5) Purge haplotigs, using any program. Use the diagnostic contigs of known dodgy origin/quality to see if they get correctly filtered out
6) Plot the ONT and S. vul onto the PB original assembly to scaffold any final contigs that need it.

with mapping - use the --sorted flag and also sort the bam files.

CURATION STEP 1: manual breaking of fasta sequences

SAM to BAM for reinvestigating using IGV for breakpoints

```
#!/bin/bash -e

#SBATCH --job-name=2023_01_04.polished_genome_sam_bam.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=10GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

module purge
module load SAMtools/1.13-GCC-9.2.0

DIR=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_dec/Basecalling_guppy6.2.1/nextpolish

GENOME_STEP0=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_dec/Basecalling_guppy6.2.1/nextpolish/Rd2_ONT_no_alts_scaffs_nextpolish_rd1.rmdup.fa
```

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step1_splitsequences
cp $GENOME_STEP0 Atri_polished_final.fasta

samtools view -S -b $DIR/ONT_reads_aligned_to_polished_241022.sam | samtools sort  > ONT_reads_aligned_to_polished_241022.bam
samtools index ONT_reads_aligned_to_polished_241022.bam
samtools view -S -b $DIR/ontPB__aligned_to_polished_241022.sam | samtools sort  > ontPB__aligned_to_polished_241022.bam
samtools index ontPB__aligned_to_polished_241022.bam
samtools view -S -b $DIR/Stuvul_aligned_to_polished_241022.sam | samtools sort  > Stuvul_aligned_to_polished_241022.bam
samtools index Stuvul_aligned_to_polished_241022.bam
samtools view -S -b $DIR/Zf_aligned_to_polished_241022.sam | samtools sort  > Zf_aligned_to_polished_241022.bam
samtools index Zf_aligned_to_polished_241022.bam

samtools faidx Atri_polished_final.fasta
```

Manually break contigs

use getfasta to create new broken contigs: https://bedtools.readthedocs.io/en/latest/content/tools/getfasta.html

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step1_splitsequences
module load SAMtools/1.13-GCC-9.2.0
module load BEDTools/2.30.0-GCC-11.3.0
samtools faidx Atri_polished_final.fasta
cut -f1,2 Atri_polished_final.fasta.fai > sizes.genome
awk '{print $1"\t"0"\t"$2}' sizes.genome > sizes.genome.bed
#file sequences_to_break_round1.bed created manually during IGV inspection/curation
cut -f1 sequences_to_break_round1.bed | tail -n +2 | sort | uniq > sequences_to_break_round1_sequencenames.txt
cat <(grep -v -f sequences_to_break_round1_sequencenames.txt sizes.genome.bed) <(tail -n +2 sequences_to_break_round1.bed) > sequences_to_break_round1_completegenome.bed
```

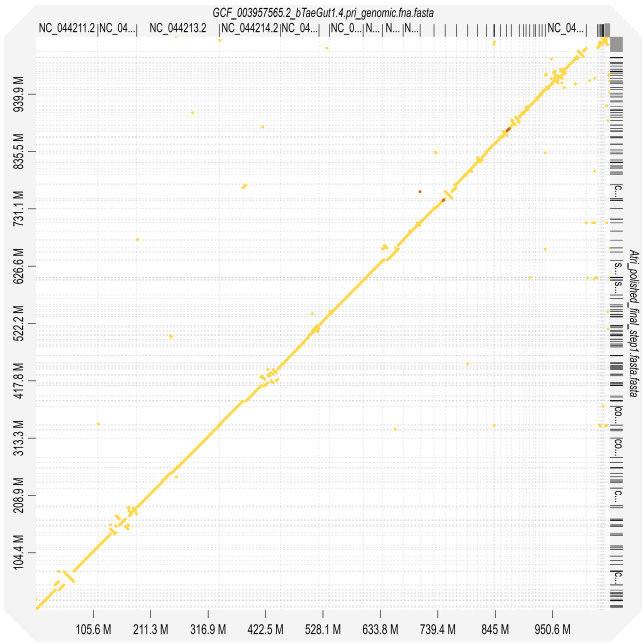
```
bedtools getfasta -fi Atri_polished_final.fasta -bed sequences_to_break_round1_completegenome.bed > Atri_polished_final_step1.fasta
#checking genome size the same, just broken up
grep -v "^>" Atri_polished_final.fasta | wc -c #1045897693
bedtools getfasta -fi Atri_polished_final.fasta -bed sequences_to_break_round1_completegenome.bed | grep -v "^>" | wc -c #1045897693

module load Python/2.7.14-gimkl-2017a
python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/tools/seqsuite.py -seqin Atri_polished_final.fasta -summarise -dna
python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/tools/seqsuite.py -seqin Atri_polished_final_step1.fasta -summarise -dna
```

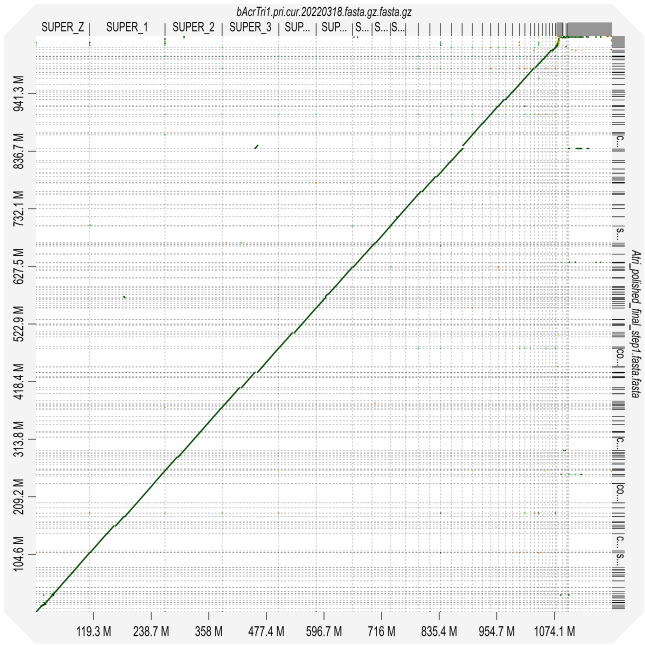
```
#~~#    00:00:03        # ~~~~~~~ Sequence Summary for Atri_polished_final ~~~~~~~ #
#SUM    00:00:32        Total number of sequences: 1,648
#SUM    00:00:32        Total length of sequences: 1,045,896,045
#SUM    00:00:32        Min. length of sequences: 13
#SUM    00:00:32        Max. length of sequences: 35,044,984
#SUM    00:00:32        Mean length of sequences: 634,645.66
#SUM    00:00:32        Median length of sequences: 2,252
#SUM    00:00:32        N50 length of sequences: 11,263,658
#SUM    00:00:32        L50 count of sequences: 29

#~~#    00:00:04        # ~~~~ Sequence Summary for Atri_polished_final_step1 ~~~~ #
#SUM    00:00:42        Total number of sequences: 1,740
#SUM    00:00:42        Total length of sequences: 1,045,895,953
#SUM    00:00:42        Min. length of sequences: 13
#SUM    00:00:42        Max. length of sequences: 35,044,984
#SUM    00:00:42        Mean length of sequences: 601,089.63
#SUM    00:00:42        Median length of sequences: 2,507
#SUM    00:00:42        N50 length of sequences: 10,406,399
#SUM    00:00:42        L50 count of sequences: 30
```

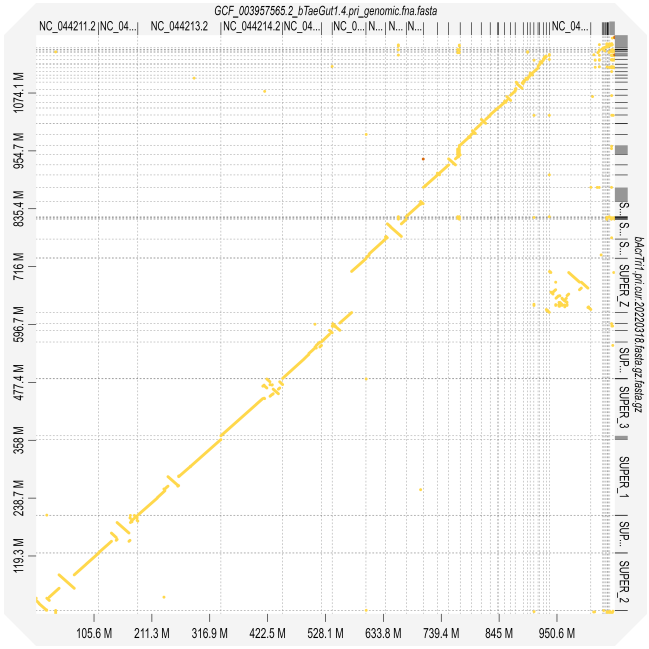
CURATION STEP 2: revisualise in dgenies

https://dgenies.toulouse.inra.fr/run

```
#!/bin/bash -e

#SBATCH --job-name=2023_01_05.Minimap2_paf_polished_step2.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-10:00:00
#SBATCH --mem=8GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=16
#SBATCH --profile task

#load modules
module load minimap2/2.24-GCC-9.2.0

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step2_dgenies
REF=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/GCF_003957565.2_bTaeGut1.4.pri/GCF_003957565.2_bTaeGut1.4.pri_genomic.fna
INDEX=/nesi/nobackup/uoa02613/kstuart_projects/programs/dgenies/index.py
GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step1_splitsequences/Atri_polished_final_step1.fasta
GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step1_splitsequences/Atri_polished_final.fasta

python ${INDEX} -i ${GENOME} -n $(basename $GENOME.fasta) -o $(basename $GENOME.fasta).idx
minimap2 -x asm5 ${REF} ${GENOME} > tgut_$(basename $GENOME .fasta).paf

REF2=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/Sturnus_vulgaris_2.3.1.simp.fasta

minimap2 -x asm5 ${REF2} ${GENOME} > svul_$(basename $GENOME .fasta).paf

REF3=/nesi/nobackup/uoa02613/Ref_genomes/bAcrTri1.pri.cur.20220318.fasta.gz
REF3=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/myna_genome_versions/VGP_jan/ncbi_dataset/data/GCA_027559615.1/GCA_027559615.1_bAcrTri1.pri_genom

python ${INDEX} -i ${GENOME} -n $(basename $GENOME.fasta) -o $(basename $GENOME.fasta).idx
python ${INDEX} -i ${REF3} -n $(basename $REF3.fasta.gz) -o $(basename $REF3.fasta.gz).idx
minimap2 -x asm5 ${REF3} ${GENOME} > AtrisVGP_$(basename $GENOME .fasta).paf

minimap2 -x asm5 ${REF} ${REF3} > tgut_AtrisVGP.paf

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step1_splitsequences/Atri_polished_final.fasta
python ${INDEX} -i ${GENOME} -n $(basename $GENOME.fasta) -o $(basename $GENOME.fasta).idx
minimap2 -x asm5 ${REF3} ${GENOME} > AtrisVGP_$(basename $GENOME .fasta).paf
```







Iterated step 1 and 2 as needed until I was happy all the missassemblies had been patched.

Check how the manual curation has impacted BUSCO score

```
#!/bin/bash -e

#SBATCH --job-name=2023_01_07.busco_step2.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-24:00:00
#SBATCH --mem=15GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=8
#SBATCH --profile task

#load modules
module purge
module load BUSCO/5.3.2-gimkl-2020a

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step2_dgenies

busco -c 8 -i ../step1_splitsequences/Atri_polished_final_step1.fasta -l aves_odb10 -m genome -o Atri_polished_final_step1
```

C:97.2%[S:96.7%,D:0.5%],F:0.5%,M:2.3%,n:8338

8110 Complete BUSCOs (C)

8067 Complete and single-copy BUSCOs (S)

43 Complete and duplicated BUSCOs (D)

45 Fragmented BUSCOs (F)

183 Missing BUSCOs (M)

8338 Total BUSCO groups searched

Below was run to investigate BUSCOs lost in this step and step 3:

Working out which contigs the now missing BUSCOs are located on

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step2_dgenies
STEP0=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_dec/Basecalling_guppy6.2.1/myna_final_polished_busco/run_aves_odb10/full_table.tsv
STEP2=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step2_dgenies/Atri_polished_final_step1/run_aves_odb10/full_table.tsv
```

```
comm -13 <(grep "Missing" $STEP0 | sort) <(grep "Missing" $STEP2 | sort) > missing_buscos.txt
awk 'FNR==NR{a[$1];next} (($1) in a)' missing_buscos.txt $STEP0
```

```
20567at8782	Complete	contig_802_np11	5739747 5744068
23551at8782	Complete	contig_1514_np11	1739491 1796678
24501at8782	Complete	contig_1290_np11	2547187 2591588
47289at8782	Complete	contig_1571_np11	8916464 8934038
```

BUSCOMP: create the symbolic links needed

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step2_dgenies/buscomp/runs

ln -s /nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_dec/Basecalling_guppy6.2.1/myna_final_polished_busco/ myna_final_polished
ln -s /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step2_dgenies/Atri_polished_final_step1/ .
ln -s /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/diploidocus/purge_Atri_polished_final_step3/Atri_polished_final_step3/ .

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step2_dgenies/buscomp/fastas

ln -s /nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_dec/Basecalling_guppy6.2.1/nextpolish/Rd2_ONT_no_alts_scaffs_nextpolish_rd1.rmdup.fa myna_final_polished.fasta
ln -s /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step1_splitsequences/Atri_polished_final_step1.fasta .
ln -
s  /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/diploidocus/purge_Atri_polished_final_step3/Atri_polished_final_step3.purge.fasta
```

run buscomp

```
#!/bin/bash -e

#SBATCH --job-name=2023_01_12.buscomp_step2.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=5GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step2_dgenies/buscomp

module load Python/2.7.14-gimkl-2017a
module load minimap2/2.24-GCC-11.3.0
python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/tools/buscomp.py runs="./runs/*" fastadir=./fastas genomesize=1000e6 forks=4 basefile=buscomp_run1 endextend=0
```

buscomp concluded that there were no sequences lost in this step, but maybe a few in step 3, lower down.

CURATION STEP 3: clean up sequences

get univec database:

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/contaminants
wget -r ftp.ncbi.nlm.nih.gov/pub/UniVec/UniVec
```

Diploidocus

```
#!/bin/bash -e

#SBATCH --job-name=2023_01_06.Diploidocus.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-24:00:00
#SBATCH --mem=50GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
```

```
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=3
#SBATCH --profile task

#load modules
#module load bbmap blast+ kat minimap2 purge_haplotigs samtools java/8u45 python/3.7.3
#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
module purge
module load BBMap/39.01-GCC-11.3.0
module load BLAST/2.13.0-GCC-11.3.0
module load GCC/7.4.0
module load XZ/5.2.4-GCCcore-7.4.0
module load KAT/2.4.2-gimkl-2018b-Python-3.7.3
module load minimap2/2.24-GCC-11.3.0
module load SAMtools/1.15.1-GCC-11.3.0
module load Java/11.0.4
#module load Python/3.7.3-gimkl-2018b
module load Perl/5.28.1-gimkl-2018b
module load BEDTools/2.28.0-gimkl-2018b
module load R/3.5.3-gimkl-2018b
module load Python/2.7.16-gimkl-2018b
module load purge_haplotigs/1.1.2-gimkl-2022a-Perl-5.34.1

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/diploidocus

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step1_splitsequences/Atri_polished_final_step1.fasta
BUSCO=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step2_dgenies/Atri_polished_final_step1/run_aves_odb10/full_table.tsv

READS=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_dec/Basecalling_guppy6.2.1/Read_Datasets/ONT_all_pc_raw.fa
KMERREADS="/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_dec/Basecalling_guppy6.2.1/pilon_polishing/Myna_10x_processed_R*_val_*.fq.gz"
SCREENDB=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/contaminants/UniVec
PREFIX=Atri_polished_final_step3

#export RSTUDIO_PANDOC=/Applications/RStudio.app/Contents/MacOS/pandoc #from Katana

#python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/tools/diploidocus.py -seqin $GENOME -runmode purgehap -basefile $PREFIX -busco $BUSCO -reads $READS km
10xtrim=T 10xtrim -forks 16 -screendb $SCREENDB pretrim=T

python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/tools/diploidocus.py -seqin $GENOME -runmode dipcycle -purgemode nala -basefile $PREFIX -busco $BUSCO -reac
kmerreads=\"$KMERREADS\" 10xtrim=T 10xtrim -forks 6 -screendb $SCREENDB pretrim=T
```

kicking up fuss at rscript depth calculations. trying to run this line manually interactive node. 64 gb. Ran in approx 1 hr and then the above script was restart

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/diploidocus
module load R/3.5.3-gimkl-2018b
Rscript /scale_wlg_nobackup/filesets/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/libraries/r/depthcopy.R pngdir=Atri_polished_final_step3.plots depfile=Atri_polished_final_ste
busco=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step2_dgenies/Atri_polished_final_step1/run_aves_odb10/full_table.tsv adjust=12 basefile=Atri_polish
```

Vecscreen

```
#!/bin/bash -e

#SBATCH --job-name=2023_01_16.Vecscreen.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-24:00:00
#SBATCH --mem=50GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=3
#SBATCH --profile task

module purge
module load Python/2.7.16-gimkl-2018b
module load BLAST/2.13.0-GCC-11.3.0

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/vecscreen
```

```
SCREENDB=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/contaminants/UniVec
GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/diploidocus/purge_Atri_polished_final_step3/Atri_polished_final_step3.pu
python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/tools/diploidocus.py runmode=vecscreen screendb=$SCREENDB screenmode=purge basefile=Atri_polished_final_s
vecmask=27 forks=3 keepnames=T

python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/tools/seqsuite.py -seqin Atri_polished_final_step3.purge.vecscreen.fasta -summarise -dna
```

soft masked approx. 20 sequences, length 20-100 bp.

```
#~~#    00:00:00        # ~~~~ Sequence Summary for Atri_polished_final_step3.purge.vecscreen ~~~~ #
#SUM    00:00:37        Total number of sequences: 804
#SUM    00:00:37        Total length of sequences: 1,040,700,499
#SUM    00:00:37        Min. length of sequences: 283
#SUM    00:00:37        Max. length of sequences: 35,044,984
#SUM    00:00:37        Mean length of sequences: 1,294,403.61
#SUM    00:00:37        Median length of sequences: 5,593
#SUM    00:00:37        N50 length of sequences: 10,406,399
#SUM    00:00:37        L50 count of sequences: 30
```

run purge_dups

https://github.com/dfguan/purge_dups

```
#!/bin/bash -e

#SBATCH --job-name=2023_01_16.Purge_dups_step3.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=20GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=8
#SBATCH --profile task

#load modules
module purge
module load purge_dups/1.2.6-gimkl-2022a-Python-3.10.5

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/purge_dups

KMERREADS="/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_dec/Basecalling_guppy6.2.1/pilon_polishing/Myna_10x_processed_R*_val_*.fq.gz"
echo $KMERREADS | sed 's/ /\n/g' > 10x.fofn

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/vecscreen/Atri_polished_final_step3.purge.vecscreen.fasta

READS=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_dec/Basecalling_guppy6.2.1/Read_Datasets/ONT_all_pc_raw.fa
echo $READS | sed 's/ /\n/g' >  reads.fofn
#READS=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_dec/Basecalling_guppy6.2.1/Read_Datasets/Pacbio_all_corrected.fasta
#echo $READS | sed 's/ /\n/g' >  reads_pb.fofn

pd_config.py -l purge_dup_PB_fofn -s 10x.fofn -n Atri_polished_final_step3.json $GENOME reads.fofn

DIR=/opt/nesi/CS400_centos7_bdw/purge_dups/1.2.6-gimkl-2022a-Python-3.10.5/bin

run_purge_dups.py -p bash Atri_polished_final_step3.json $DIR Atri_polished_final_step3
```

checked with both ONT and PacBio reads. No dups found!

NumtFinder

and remove the mitochondrial genome

https://www.ncbi.nlm.nih.gov/nuccore/CM050619.1

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/mitochondrial_genomes
```

Run numtfinder

```
#!/bin/bash -e

#SBATCH --job-name=2023_01_16.Numtfinder.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-10:00:00
#SBATCH --mem=5GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/numtfinder

module purge
module load Python/2.7.14-gimkl-2017a
module load BLAST/2.10.0-GCC-9.2.0

MITO=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/mitochondrial_genomes/A_tris_VGP_mito_genome_CM050619.1.fasta
#GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/vecscreen/Atri_polished_final_step3.purge.vecscreen.fasta
#GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step1_splitsequences/Atri_polished_final.fasta
GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step1_splitsequences/Atri_polished_final_step1.fasta

python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/dev/numtfinder.py seqin=${GENOME} mtdna=${MITO} basefile=$(basename $GENOME .fasta)_$(basename $MITO

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/diploidocus/purge_Atri_polished_final_step3/Atri_polished_final_step3.pu
python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/dev/numtfinder.py seqin=${GENOME} mtdna=${MITO} basefile=$(basename $GENOME .fasta)_$(basename $MITO
```

Actually turns out that mito genome was chucked out as 'JUNK' by diploidocus (contig_1670_np11).

Remove Small Sequences

```
#!/bin/bash -e

#SBATCH --job-name=2023_01_16.Trim.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-01:00:00
#SBATCH --mem=1GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

module purge
module load BBMap/38.81-gimkl-2020a

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/trim

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/vecscreen/Atri_polished_final_step3.purge.vecscreen.fasta

#remove contig tail, with min length 1,000 bp
reformat.sh in=$GENOME out=Atri_polished_final_step3.purge.vecscreen.trimmed.fasta minlength=1000

module purge
module load Python/2.7.16-gimkl-2018b

python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/tools/seqsuite.py -seqin Atri_polished_final_step3.purge.vecscreen.trimmed.fasta -summarise -dna
```

```
#~~#    00:02:34        # ~~~~ Sequence Summary for Atri_polished_final_step3.purge.vecscreen.trimmed ~~~~ #
#SUM    00:03:51        Total number of sequences: 605
#SUM    00:03:51        Total length of sequences: 1,040,569,522
#SUM    00:03:51        Min. length of sequences: 1,010
#SUM    00:03:51        Max. length of sequences: 35,044,984
#SUM    00:03:51        Mean length of sequences: 1,719,949.62
#SUM    00:03:51        Median length of sequences: 30,599
#SUM    00:03:51        N50 length of sequences: 10,406,399
#SUM    00:03:51        L50 count of sequences: 30
```

BBmap seems to encode the fasta file worse. So pulling out the contigs that were excluded (as identified by BBmap) and manually removing them from the

```
module load SAMtools/1.13-GCC-9.2.0
module load BEDTools/2.30.0-GCC-11.3.0
samtools faidx Atri_polished_final_step3.purge.vecscreen.trimmed.fasta
cut -f1,2 Atri_polished_final_step3.purge.vecscreen.trimmed.fasta.fai > sizes.genome
awk '{print $1"\t"0"\t"$2}' sizes.genome > sizes.genome.bed

bedtools getfasta -fi $GENOME -bed sizes.genome.bed  > Atri_polished_final_step3.purge.vecscreen.trim.fasta
```

Blobtools

https://blobtools.readme.io/docs/my-first-blobplot

https://blobtoolkit.genomehubs.org/install/#databases

https://blobtoolkit.genomehubs.org/blobtools2/blobtools2-tutorials/adding-data-to-a-dataset/adding-hits/

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs/blobtools
conda create -n blobtools
conda activate blobtools
conda install -c anaconda matplotlib docopt tqdm wget pyyaml git
conda install -c bioconda pysam --update-deps
```

Add databases

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs/blobtools/databases
mkdir -p taxdump;
cd taxdump;
mkdir -p taxdump;
cd taxdump;
curl -L ftp://ftp.ncbi.nih.gov/pub/taxonomy/new_taxdump/new_taxdump.tar.gz | tar xzf -;
cd ..;
```

Making bam file

```
#!/bin/bash -e

#SBATCH --job-name=2023_01_18.blobtools_mapping.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-24:00:00
#SBATCH --mem=12GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=8
#SBATCH --profile task

module purge
module load minimap2/2.24-GCC-11.3.0
module load SAMtools/1.13-GCC-9.2.0

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/blobtools

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/trim/Atri_polished_final_step3.purge.vecscreen.trim.fasta
ONT_READS=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_dec/Basecalling_guppy6.2.1/Read_Datasets/ONT_all_pc_raw.fa
```

```
#align raw ONT reads
minimap2 -ax map-ont -t 4  $GENOME $ONT_READS > ONT_reads_aligned_to_Atris_trim.sam

samtools view -S -b ONT_reads_aligned_to_Atris_trim.sam | samtools sort  > ONT_reads_aligned_to_Atris_trim.bam
samtools index ONT_reads_aligned_to_Atris_trim.bam
```

Making databases

blastn
https://github.com/blobtoolkit/pipeline/blob/master/rules/run_blastn.smk

```
#!/bin/bash -e

#SBATCH --job-name=2023_01_18.blobtools_blastn.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-24:00:00
#SBATCH --mem=20GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=16
#SBATCH --profile task

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/blobtools/blastn2

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/trim/Atri_polished_final_step3.purge.vecscreen.trim.fasta

module load BLAST/2.13.0-GCC-11.3.0
module load BLASTDB/2021-05

QUERIES=$GENOME
FORMAT="6 qseqid staxids bitscore std"
BLASTOPTS="-task megablast"
BLASTAPP=blastn
DB=nt

# Keep the database in RAM
#cp $BLASTDB/{$DB,taxdb}* $TMPDIR/
#export BLASTDB=$TMPDIR

$BLASTAPP $BLASTOPTS -db $DB -query $QUERIES -outfmt "$FORMAT" \
   -out $(basename $QUERIES .fasta).$DB.$BLASTAPP -num_threads $SLURM_CPUS_PER_TASK
```

And run blobtools

```
#!/bin/bash -e

#SBATCH --job-name=2023_01_24.blobtools_create.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-02:00:00
#SBATCH --mem=12GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=8
#SBATCH --profile task

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/blobtools

DIR=/nesi/nobackup/uoa02613/kstuart_projects/programs/blobtools/

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/trim/Atri_polished_final_step3.purge.vecscreen.trim.fasta

BLASTN=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/blobtools/blastn/Atri_polished_final_step3.purge.vecscreen.trim.nt.blastn

#Creating a blobDB
$DIR/blobtools create \
```

```
-i $GENOME \
-b ONT_reads_aligned_to_Atris_trim.bam \
-t $BLASTN \
--nodes $DIR/databases/taxdump/nodes.dmp \
--names $DIR/databases/taxdump/names.dmp \
-o my_first_blobplot


#Creating a blobplot
$DIR/blobtools plot \
-i my_first_blobplot.blobDB.json \
--format pdf, svg, tiff

#Filtering the fasta
$DIR/blobtools view \
-i my_first_blobplot.blobDB.json \
```

Apicomplexa:

Actinomycetota:

Streptophyta:

Euglenozoa:

investigating which of the sequences need removing post blobtooling

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/blobtools
grep -v "Chordata\|no-hit" my_first_blobplot.blobDB.table.txt | cut -f1 | grep -v "^#" > blobtools_contaminated_sequences.txt

DIR=/nesi/nobackup/uoa02613/kstuart_projects/programs/blobtools/
GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/trim/Atri_polished_final_step3.purge.vecscreen.trim.fasta
$DIR/blobtools seqfilter -v \
-i $GENOME \
-l  blobtools_contaminated_sequences.txt

module purge
module load Python/2.7.16-gimkl-2018b

python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/tools/seqsuite.py -seqin Atri_polished_final_step3.purge.vecscreen.trim.filtered.fna -summarise -dna
```

CURATION STEP 4: synteny scaffolding

RagTag:

https://github.com/malonge/RagTag/wiki/scaffold

installation

```
cd /nesi/nobackup/uoa02613/kstuart_projects/programs
echo "export PATH=\$PATH:\/nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/bin" >> $HOME/.bashrc # add to .bashrc
source $HOME/.bashrc
conda init
conda install -c bioconda ragtag
conda create -n ragtag tagtag#didn't work
#ragtag .py's here for some reason? Not sure why they installed as single executables and not as an env.
/nesi/nobackup/uoa02613/kstuart_projects/programs/miniconda/bin
```

Working with the ONT_noalt_scaff assembly

```
#!/bin/bash -e

#SBATCH --job-name=2023_01_24.step4_scaffold.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-02:00:00
#SBATCH --mem=12GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
```

```
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding

REF_ATRIS=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/myna_genome_versions/VGP_jan/ncbi_dataset/data/GCA_027559615.1/GCA_027559615.1_bAcrTri1.pri_

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/blobtools/Atri_polished_final_step3.purge.vecscreen.trim.filtered.fna

ragtag.py scaffold ${REF_ATRIS} ${GENOME} -t 8 -o ./ragtag_atris_synteny -j scaff_1541.txt

module purge
module load Python/2.7.16-gimkl-2018b

cd ragtag_atris_synteny
python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/tools/seqsuite.py -seqin ragtag.scaffold.fasta -summarise -dna
```

renaming sequence names

```
#!/bin/bash -e

#SBATCH --job-name=2023_01_24.Minimap2_renaming.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-10:00:00
#SBATCH --mem=8GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=16
#SBATCH --profile task

#load modules
module load minimap2/2.24-GCC-9.2.0

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/renamed

INDEX=/nesi/nobackup/uoa02613/kstuart_projects/programs/dgenies/index.py
REF=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/GCF_003957565.2_bTaeGut1.4.pri/GCF_003957565.2_bTaeGut1.4.pri_genomic.fna
GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/ragtag.scaffold.fasta

python ${INDEX} -i ${GENOME} -n $(basename $GENOME.fasta) -o $(basename $GENOME.fasta).idx
minimap2 -x asm5 ${REF} ${GENOME} > tgut_$(basename $GENOME .fasta).paf
```

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/renamed

#used below to help with my naming scheme
grep "^>" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/myna_genome_versions/VGP_jan/ncbi_dataset/data/GCA_027559615.1/GCA_027559615.1_bAcrTri1.pri_genc
VGPmyna_sequence_names.txt
grep "^>" ../ragtag.scaffold.fasta > myna_sequence_names.txt
grep "^>" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/GCF_003957565.2_bTaeGut1.4.pri/GCF_003957565.2_bTaeGut1.4.pri_genomic.fna > zebrafinch_sequence_n

#manually sorted out a reasonable scaffold naming scheme
module load SeqKit/2.2.0
awk -F "\t" '{print $1,"\t",$6}' renaming.txt | sed 's/>\|"\| //g' > renaming_scaffoldnames.txt
seqkit replace -p "(.+)" -r '{kv}' -k renaming_scaffoldnames.txt ../ragtag.scaffold.fasta > AcTris_vAus2.0_unordered.fasta

#check it worked
comm -12 <(grep "^>" AcTris_vAus2.0_unordered.fasta | sed 's/>//g' | sort) <(cut -f2 renaming_scaffoldnames.txt | sort) | wc -l

#reorder the sequences
module load SAMtools/1.15.1-GCC-11.3.0
cut -f2 renaming_scaffoldnames.txt | tail -n +2 > renaming_scaffoldnames_order.txt
samtools faidx AcTris_vAus2.0_unordered.fasta
samtools faidx AcTris_vAus2.0_unordered.fasta  $(cat renaming_scaffoldnames_order.txt) | sed 's/k/ /g' > AcTris_vAus2.0.fasta
```

```
#~~#    00:00:00       # ~~~~~~~~~ Sequence Summary for ragtag.scaffold ~~~~~~~~~ #
#SUM    00:00:56       Total number of sequences: 256
#SUM    00:00:56       Total length of sequences: 1,040,539,946
#SUM    00:00:56       Min. length of sequences: 1,010
#SUM    00:00:56       Max. length of sequences: 150,861,042
#SUM    00:00:56       Mean length of sequences: 4,064,609.16
#SUM    00:00:56       Median length of sequences: 3,369
#SUM    00:00:56       N50 length of sequences: 72,486,765
#SUM    00:00:56       L50 count of sequences: 5
#SUM    00:00:56       Total number of contigs: 597
#SUM    00:00:56       Contig N50 length of sequences: 10,406,399
#SUM    00:00:56       Contig L50 count of sequences: 30
#SUM    00:00:56       GC content: 41.85%
#SUM    00:00:56       N bases: 34,958 (0.00%)
#SUM    00:00:56       Gap (10+ N) length: 34,100 (0.00%)
#SUM    00:00:56       Gap (10+ N) count: 341
```

UPDATED: Renaming the W chrom fragment (incorrectly labeled as W chrom, is unplaced scaffold fragment)

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/renamed

#edit the name of the scaffold and the order file (latter done manually)
sed 's/Superscaffold_chrW/Scaffold_unplaced_13/g' renaming_scaffoldnames.txt > renaming_scaffoldnames_updated.txt

#manually sorted out a reasonable scaffold naming scheme
module load SeqKit/2.2.0
seqkit replace -p "(.+)" -r '{kv}' -k renaming_scaffoldnames_updated.txt ../ragtag.scaffold.fasta > AcTris_vAus2.1_unordered.fasta

#check it worked
comm -12 <(grep "^>" AcTris_vAus2.1_unordered.fasta | sed 's/>//g' | sort) <(cut -f2 renaming_scaffoldnames_updated.txt | sort) | wc -l

#reorder the sequences
module load SAMtools/1.15.1-GCC-11.3.0
samtools faidx AcTris_vAus2.1_unordered.fasta
samtools faidx AcTris_vAus2.1_unordered.fasta  $(cat renaming_scaffoldnames_order_updated.txt) | sed 's/k/ /g' > AcTris_vAus2.1.fasta
```

NCBI: Various upload checks

```
module purge
module load Python/2.7.16-gimkl-2018b

python /nesi/nobackup/uoa02613/kstuart_projects/programs/SLiMSuite/tools/seqsuite.py -seqin ONT_all_pc_raw.fa -summarise -dna
```

```
#~~#    00:03:56       # ~~~~~~~~~ Sequence Summary for ONT_all_pc_raw ~~~~~~~~~~ #
#SUM    00:21:09       Total number of sequences: 3,324,660
#SUM    00:21:09       Total length of sequences: 23,552,201,852
#SUM    00:21:13       Min. length of sequences: 13
#SUM    00:21:13       Max. length of sequences: 142,138
#SUM    00:21:13       Mean length of sequences: 7,084.09
#SUM    00:21:13       Median length of sequences: 4,987
#SUM    00:21:13       N50 length of sequences: 11,866
#SUM    00:21:13       L50 count of sequences: 616,368
#SUM    00:21:13       Total number of contigs: 3,324,660
#SUM    00:21:13       GC content: 43.37%
```

```
#SUM    00:21:13    N bases: 0 (0.00%)
#SUM    00:21:13    Gap (10+ N) length: 0 (0.00%)
#SUM    00:21:13    Gap (10+ N) count: 0
#RUN    00:21:13    SeqList V1.48.0 run finished.
```

Looking at Z chrom mystery

using linkage

Determine which scaffs/contigs map to the three groupings: Moving scaffold: scaffold_1541_np11 (from dgenies plot)

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step2_dgenies/linkage
grep "scaffold_1541_np11" $ZF_PAF  | cut -f1 | sort | uniq > atris_scaff1541.txt
ZF_PAF=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step2_dgenies/tgut_Atri_polished_final_step1.paf
#find Atri_polished_final_step1.fasta contigs that map to zebra finch chrom 5 NC_044217.2, excluding 1541
grep "NC_044217.2" $ZF_PAF  | grep -v "scaffold_1541_np11" | cut -f1 | sort | uniq > tugt_atris_chrom5.txt
AT_PAF=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step2_dgenies/AtrisVGP_Atri_polished_final_step1.paf
#find Atri_polished_final_step1.fasta contigs that map to chrom z SUPER_Z, excluding 1541
grep "SUPER_Z" $AT_PAF  | grep -v "scaffold_1541_np11" | cut -f1 | sort | uniq > atrisVGP_atris_chromZ.txt

#rename the OLDatris assembly so that the contig names match the vcf file contig names
sed 's/>/>contig_/g' myna_s2.1.fasta > myna_s2.1_renamed.fasta
```

Figure out which of the old myna assembly's contigs match to these lists

```
#!/bin/bash -e

#SBATCH --job-name=2023_01_13.zchrom_linkage.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=15GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

module load minimap2/2.24-GCC-9.2.0

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step2_dgenies/linkage

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step1_splitsequences/Atri_polished_final_step1.fasta
minimap2 -x asm5 ${GENOME} myna_s2.1.fasta > mynas2.1_Atristep1.paf
```

Determine which scaffs/contigs map to the three groupings: Moving scaffold: scaffold_1541_np11 (from dgenies plot)

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step2_dgenies/linkage
PAF=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step2_dgenies/linkage/mynas2.1_Atristep1.paf
cat mynas2.1_Atristep1.paf | awk '$11>5000000' > mynas2.1_Atristep1_curated.paf

grep -f atris_scaff1541.txt mynas2.1_Atristep1_curated.paf  | cut -f1 | sort | uniq > OLDatris_scaff1541.txt
grep -f tugt_atris_chrom5.txt mynas2.1_Atristep1_curated.paf  | cut -f1 | sort | uniq > OLDatris_chrom5.txt
grep -f atrisVGP_atris_chromZ.txt mynas2.1_Atristep1_curated.paf  | cut -f1 | sort | uniq > OLDatris_chromz.txt

#gram unique chrom5 and chromz contigs
comm -23 OLDatris_chrom5.txt   <(cat OLDatris_scaff1541.txt OLDatris_chromz.txt | sort) > OLDatris_chrom5_uniq.txt
comm -23 OLDatris_chromz.txt   <(cat OLDatris_scaff1541.txt OLDatris_chrom5.txt | sort) > OLDatris_chromz_uniq.txt

#concat all these three groups of contigs together
cat OLDatris_scaff1541.txt  OLDatris_chrom5_uniq.txt  OLDatris_chromz_uniq.txt > OLDatris_contigfile.txt

wc -l OLDatris_contigfile.txt
#actually played with the column 11 cutoff to find the biggest scaffolds, 1 representative from each of the three files, and am using this
chrom 5: 59
chrom z: 51
#redo on monday with more chroms per test ones, and also with higher MAF threshold
```

calculate r2 values of this subset

```
#!/bin/bash -e

#SBATCH --job-name=2023_01_13.zchrom_linkageR2.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-02:00:00
#SBATCH --mem=1GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

module load VCFtools/0.1.15-GCC-9.2.0-Perl-5.30.1
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/zchrom_mystery/linkage
vcftools --vcf populations.ALL.bialminGQ30DP15-125.norep.noadm.highnegfis.lmiss20.nosingledoubletons.chr_renamed.snps.vcf --maf 0.1  --chr contig_59 --chr contig_51 --chr contig_130
```

only grab rows with values of the comparison region

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/zchrom_mystery/linkage

grep "contig_130980" out.interchrom.geno.ld | grep  "contig_51" | cut -f6 | grep -v "nan" | datamash mean 1 sstdev 1
grep "contig_130980" out.interchrom.geno.ld | grep  "contig_59" | cut -f6 | grep -v "nan" | datamash mean 1 sstdev 1

awk '$6 > 0.05' out.interchrom.geno.ld | grep -v "nan"  > out.interchrom.geno.ld_highthresh

grep "contig_130980" out.interchrom.geno.ld | awk '$6 > 0.05' | grep "contig_51" | wc -l #
grep "contig_130980" out.interchrom.geno.ld | grep "contig_51" | wc -l #
grep "contig_130980" out.interchrom.geno.ld | awk '$6 > 0.05' | grep "contig_59" | wc -l #
grep "contig_130980" out.interchrom.geno.ld | grep "contig_59" | wc -l #
```

using mapping of ONT reads

```
#!/bin/bash -e

#SBATCH --job-name=2023_01_16.zchrom_mapping.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=20GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=4
#SBATCH --profile task

module purge
```

```
module load minimap2/2.24-GCC-11.3.0
module load SAMtools/1.13-GCC-9.2.0

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/zchrom_mystery

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/bAcrTri1.pri.cur.20220318.fasta

#grep -n "^>" $GENOME | head -n 2 #1851189
#head -n 1851188 $GENOME > bAcrTri1.pri.cur.20220318.zchrom.fasta

ONT_REDS=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_dec/Basecalling_guppy6.2.1/Read_Datasets/ONT_all_pc_raw.fa

#align raw ONT reads
minimap2 -ax map-ont -t 4 bAcrTri1.pri.cur.20220318.zchrom.fasta  $ONT_REDS > ONT_reads_aligned_to_AtrisVGP_zchrom.sam

samtools view -S -b ONT_reads_aligned_to_AtrisVGP_zchrom.sam | samtools sort  > ONT_reads_aligned_to_AtrisVGP_zchrom.bam
samtools index ONT_reads_aligned_to_AtrisVGP_zchrom.bam

samtools faidx bAcrTri1.pri.cur.20220318.zchrom.fasta

#aligning zebra finch
REF=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/data/GCF_003957565.2_bTaeGut1.4.pri/GCF_003957565.2_bTaeGut1.4.pri_genomic.fna

minimap2 -ax map-ont -t 4 bAcrTri1.pri.cur.20220318.zchrom.fasta  $REF > Tgut_aligned_to_AtrisVGP_zchrom.sam

samtools view -S -b Tgut_aligned_to_AtrisVGP_zchrom.sam | samtools sort  > Tgut_aligned_to_AtrisVGP_zchrom.bam
samtools index Tgut_aligned_to_AtrisVGP_zchrom.bam

#aligning our assembly
GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/trim/Atri_polished_final_step3.purge.vecscreen.trim.fasta

minimap2 -ax map-ont -t 4 bAcrTri1.pri.cur.20220318.zchrom.fasta  $GENOME > Atris_aligned_to_AtrisVGP_zchrom.sam

samtools view -S -b Atris_aligned_to_AtrisVGP_zchrom.sam | samtools sort  > Atris_aligned_to_AtrisVGP_zchrom.bam
samtools index Atris_aligned_to_AtrisVGP_zchrom.bam
```

Looking at W chrom mystery

using linkage

Determine which scaffs/contigs map to the three groupings: Moving scaffold: scaffold_1541_np11 (from dgenies plot)

```
cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/wchrom
GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/renamed/AcTris_vAus2.0.fasta

grep -n "Superscaffold_chrW" $GENOME #15825784
grep -n "Superscaffold_chrZ" $GENOME  #15828923

grep "JAPZMO010000087.1" ragtag.scaffold.agp
```

JAPZMO010000087.1_RagTag 1 166463 1 W contig_1971_np11:0-166463:0-166463 1 166463 -

JAPZMO010000087.1_RagTag 166464 166563 2 U 100 scaffold yes align_genus #U indicates gap of 100 bp

JAPZMO010000087.1_RagTag 166564 188252 3 W contig_2102_np11:0-21689:0-21689 1 21689 -

```
GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/blobtools/Atri_polished_final_step3.purge.vecscreen.trim.filtered.fna
module load BEDTools/2.30.0-GCC-11.3.0
```

```
bedtools getfasta -fi $GENOME -bed wchrom_contigs.bed > wchrom_contigs.fasta
```

minimap

```
#!/bin/bash -e

#SBATCH --job-name=2023_02_13.wchrom_contig.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=5GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

module load minimap2/2.24-GCC-9.2.0

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/wchrom

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step3_cleansequences/blobtools/Atri_polished_final_step3.purge.vecscreen.trim.filtered.fna
minimap2 -x asm5 ${GENOME} wchrom_contigs.fasta > wchrom_contigs.paf
```

repeats

```
#!/bin/bash -e

#SBATCH --job-name=2023_02_13.wchrom_repeats.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=5GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

module load RepeatMasker/4.1.0-gimkl-2020a

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/wchrom

RepeatMasker -pa 2  -species aves -dir . wchrom_contigs.fasta
```

Basically all repeats!

For checking strange PB coverage:

Why are the pacbio reads so strange? Very uneven coverage - any reason for this?

Find regions of high ONT coverage, separate these these. Find regions of high PB coverage, and low PB coverage. When circlize plot tracks are made, which for differences in these three groups of bed inter

need to sort! to save mem

```
#!/bin/bash -e
#SBATCH --job-name=2022_12_12.Genome_PB_ONT_coverage.sl
#SBATCH --account=uoa02613
```

```
#SBATCH --time=00-12:00:00
#SBATCH --mem=50GB
#SBATCH --output=%x_%j.errout
#SBATCH --mail-user=katarina.stuart@auckland.ac.nz
#SBATCH --mail-type=ALL
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --cpus-per-task=2
#SBATCH --profile task

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/pb_coverage

#load modules
module load SAMtools/1.13-GCC-9.2.0
module load BEDTools/2.30.0-GCC-11.3.0

#define paths
GENOME=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_sept/genome_QC/Assembly_ONT_noalt_scaf_medaka.fasta
DIR=/nesi/nobackup/uoa02613/Myna_ONT_2022/Myna_ONT_RAW_sept/Basecalling_guppy6.2.1/curation
WIDTH=10000 #10kb windows seem like a reasonable diagnostic bin size when looking at IGV plots

#genome scaffold sizes, then break into windows
samtools faidx $GENOME
cut -f1,2 ${GENOME}.fai > sizes.genome
bedtools makewindows -g sizes.genome -w ${WIDTH} >Assembly_ONT_noalt_scaf_medaka_${WIDTH}bps.bed

#find coverage of BAM files for PB and ONT data
bedtools coverage -a Assembly_ONT_noalt_scaf_medaka_${WIDTH}bps.bed -b ${DIR}/corPB_reads_aligned_to_ONT_raw_noalts_scafs_241022.sorted.bam -counts > PBreads_${WIDTH}bp
bedtools coverage -a Assembly_ONT_noalt_scaf_medaka_${WIDTH}bps.bed -b ${DIR}/ONT_reads_aligned_to_ONT_raw_noalts_scafs_241022.sorted.bam -counts > ONTreads_${WIDTH}bp
```