Starling-May18

Projects/Katarina Stuart/KStuart.Starling-Aug18/At1_Genome/Analysis/2023-01-27.WGSresequencing

PDF Version generated by

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

on

Aug 23, 2023 @09:50 AM NZST

Table of Contents

2023-01-27.WGSresequencing

Katarina Stuart (z5188231@ad.unsw.edu.au) - Feb 28, 2023, 8:32 PM GMT+13

SNP var calling: Myna and AcTris_vAU2.0

Planning:

Mapping: BWA mem (on different page)

http://bio-bwa.sourceforge.net/bwa.shtml

Variant Calling: Samtools BCFtools

https://samtools.github.io/bcftools/howtos/variant-calling.html

Raw data:

```
/nesi/nobackup/uoa02613/Myna_WGS_2022/82_bird_data/Cleandata/
```

MAPPING WITH BWA

Create a BWA genome database: (completed in batch 1 analysis)

```
#!/bin/bash -e

#SBATCH --job-name=2023_01_30.VarCalling_index_myna.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-02:00:00
#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 BWA/0.7.17-GCC-9.2.0

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

bwa index $GENOME
```

Trimming with TrimGalore: #Done already by annabel previously

```
#!/bin/bash -e

#SBATCH --job-name=2023_01_30.rawdata_myna_AU_wgs_trim.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=16
#SBATCH --profile task
```

```
#SBATCH --array=1-49

module load TrimGalore/0.6.7-gimkl-2020a-Python-3.8.2-Perl-5.30.1

FILE=$(sed "${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/At4_MynaStarling/rawdata/starling_AU_wgs/sample_individual_list.txt)
SAMPLE=$(basename $FILE .1.fq.gz)
echo "working with sample:" $SAMPLE

mkdir /nesi/nobackup/uoa02613/kstuart_projects/At4_MynaStarling/rawdata/starling_AU_wgs_trimmed/${SAMPLE}

OUTPUT_DIR=/nesi/nobackup/uoa02613/kstuart_projects/At4_MynaStarling/rawdata/starling_AU_wgs_trimmed/${SAMPLE}
RAW_DATA_R1=/nesi/nobackup/uoa02613/kstuart_projects/At4_MynaStarling/rawdata/starling_AU_wgs/${SAMPLE}*1.fq.gz
RAW_DATA_R2=/nesi/nobackup/uoa02613/kstuart_projects/At4_MynaStarling/rawdata/starling_AU_wgs/${SAMPLE}*2.fq.gz

trim_galore -j 16 -o ${OUTPUT_DIR} --fastqc --paired ${RAW_DATA_R1} ${RAW_DATA_R2}
```

Aligning with bwa mem:

```
#!/bin/bash -e

#SBATCH --job-name=2022_02_28.VarCalling_myna_wgs_map.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=16
#SBATCH --profile task
#SBATCH --array=1-18

# load modules
module load BWA/0.7.17-GCC-9.2.0
module load SAMtools/1.15.1-GCC-11.3.0

cd /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/variant_calling/mapped_reads

# set paths
SAMPLE=$(sed
"${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/psmc/psmc_individuals_subset_round2.txt)
echo "working with sample:" $SAMPLE

GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/renamed/AcTris_vAus2.0.fasta
TRIM_DATA_R1=/nesi/nobackup/uoa02613/Myna_WGS_2022/82_bird_data/Cleandata/${SAMPLE}/${SAMPLE}_R1.fq.gz
TRIM_DATA_R2=/nesi/nobackup/uoa02613/Myna_WGS_2022/82_bird_data/Cleandata/${SAMPLE}/${SAMPLE}_R2.fq.gz
OUT_DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/variant_calling/mapped_reads

# Map the reads
bwa mem -t ${SLURM_CPUS_PER_TASK} \
-R "@RG\tID:${SAMPLE}\tLB:${SAMPLE}_WGS\tPL:ILLUMINA\tSM:${SAMPLE}" \
-M ${GENOME} ${TRIM_DATA_R1} ${TRIM_DATA_R2} | \
samtools sort | samtools view -O BAM -o ${OUT_DIR}/${SAMPLE}.sorted.bam

# Check output
#samtools flagstat ${OUT_DIR}/${SAMPLE}.sorted.bam
```

Mark duplicates with picard:

```
#!/bin/bash -e
#SBATCH --job-name=2023_02_28.VarCalling_myna_wgs_dup.sl
#SBATCH --account=uoa02613
```

```
#SBATCH --time=00-12:00:00
#SBATCH --mem=60GB
#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
#SBATCH --array=1-18

# load modules
module load picard/2.26.10-Java-11.0.4
module load SAMtools/1.15.1-GCC-11.3.0

# set paths
SAMPLE=$(sed
"${SLURM_ARRAY_TASK_ID}q;d" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/demography/psmc/psmc_individuals_subset_round2.txt)
echo "working with sample:" $SAMPLE

OUT_DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/variant_calling/mapped_reads

TMPDIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/variant_calling/mapped_reads/tmp

export _JAVA_OPTIONS=-Djava.io.tmpdir=${TMPDIR}

#Mark Duplicates

picard MarkDuplicates INPUT=${OUT_DIR}/${SAMPLE}.sorted.bam OUTPUT=${OUT_DIR}/${SAMPLE}.sorted.dup.bam
METRICS_FILE=${OUT_DIR}/${SAMPLE}.metrics.txt MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1000;

# Generate index

samtools index -@ ${SLURM_CPUS_PER_TASK} ${OUT_DIR}/${SAMPLE}.sorted.dup.bam
```

Variant Calling with BCFtools:

Call variants: **Run time 4-18:50:00**

```
#!/bin/bash -e

#SBATCH --job-name=2023_02_01.VarCalling_myna_wgs_mpileup.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-150:00:00
#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=2
#SBATCH --profile task

# load modules
module load BCFtools/1.13-GCC-9.2.0
module load SAMtools/1.15.1-GCC-11.3.0

#create file where BAM files list will go
#OUT_DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/variant_calling/SNP_bcftools

#touch ${OUT_DIR}/sample_bamfiles_list.txt

#Fill BAM file list. Basename command now redundant due to fixed list.
```

```
#for SAMPLE_NUMBER in {1..42}
#do
#SAMPLE=$(sed
"${SAMPLE_NUMBER}q;d" /nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/variant_calling/sample_individual_list_ALL.txt)
#DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/variant_calling/mapped_reads
#BAM=${DIR}/${SAMPLE}.sorted.dup.bam
#echo ${BAM} >> ${OUT_DIR}/sample_bamfiles_list.txt
#done

# set paths
GENOME=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/curation/step4_scaffolding/ragtag_atris_synteny/renamed/AcTris_vAus2.0.fasta
OUT_DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/variant_calling/SNP_bcftools
BAM_LIST=${OUT_DIR}/sample_bamfiles_list.txt

## Calling variants

bcftools mpileup -C 50 -q 20 -Q 25 -a 'DP,AD,ADF,ADR,SP' -Ou -f ${GENOME} -b ${BAM_LIST} | \
bcftools call -c | \
bcftools view --exclude-types indels | \
bcftools sort --temp-dir ${OUT_DIR}/temp -Oz -o ${OUT_DIR}/myna_42inds.vcf.gz
```

Some basic filtering

```
#!/bin/bash -e

#SBATCH --job-name=2023_02_07.VarCalling_myna_wgs_filtering.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=2GB
#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=1
#SBATCH --profile task

# load modules
module load BCFtools/1.13-GCC-9.2.0
module load SAMtools/1.10-GCC-9.2.0
module load VCFtools/0.1.15-GCC-9.2.0-Perl-5.30.1

# set paths
DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/variant_calling/SNP_bcftools
VCF=${DIR}/myna_42inds.vcf.gz

vcftools --gzvcf ${VCF} --max-missing 0.5 --maf 0.03 --min-meanDP 2 --max-meanDP 50 --min-alleles 2 --max-alleles 2 --recode --out ${DIR}/myna_42inds_filtered

# Discard records with r2 bigger than 0.6 in a window of 1000 sites

bcftools +prune -l 0.6 -w 1000${DIR}/myna_42inds_filtered.recode.vcf -Ov -o ${DIR}/myna_42inds_filtered_r2.vcf
```

```
#!/bin/bash -e

#SBATCH --job-name=2023_02_24.VarCalling_myna_wgs_stats.sl
#SBATCH --account=uoa02613
#SBATCH --time=00-12:00:00
#SBATCH --mem=2GB
#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=1
#SBATCH --profile task

# load modules
module load VCFtools/0.1.15-GCC-9.2.0-Perl-5.30.1

# set paths
DIR=/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/variant_calling/SNP_bcftools

cd $DIR


vcftools --vcf ${DIR}/myna_42inds_filtered.recode.vcf --het

vcftools --vcf ${DIR}/myna_42inds_filtered.recode.vcf --keep
/nesi/nobackup/uoa02613/kstuart_projects/At1_MynaGenome/analysis/psmc/psmc_individuals_subset.txt --het
```