Starling-May18
Projects/Katarina Stuart/KStuart.Starling-Aug18/Sv5_AustraliaWGS/Data/2021.05.26.SNPvariants

PDF Version generated by

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

on

Feb 20, 2023 @10:23 AM NZDT

Table of Contents

2021.05.26.SNPvariants

Katarina Stuart (z5188231@ad.unsw.edu.au) - Jun 02, 2022, 11:27 AM GMT+12

SNP Variant calling

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/genome
module load bowtie/2.3.5.1
bowtie2-build -f Sturnus_vulgaris_2.3.1.simp.fasta Sturnus_vulgaris_2.3.1.simp
```

```
#!/bin/bash
#PBS -N 2021-05-26.bam_creation.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=120gb
#PBS -l walltime=12:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae
#PBS -J 01-02

module load bowtie/2.3.5.1
module load samtools/1.10

GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/genome

DIR=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/data/snp_variants

cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/data/batch1_2_3_4_adaprem

echo "working with sample:"
sed "${PBS_ARRAY_INDEX}q;d" /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/scripts/rawdata_processing/snp_variants/sampleorder_au.txt
SAMPLE=$(sed "${PBS_ARRAY_INDEX}q;d" /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/scripts/rawdata_processing/snp_variants/sampleorder_au.txt)

bowtie2 -p 10 --phred33 --very-sensitive-local -x $GENOME/Sturnus_vulgaris_2.3.1.simp -I 149 -X 900 --rg-id ${SAMPLE} --rg SM:${SAMPLE} -1 ${SAMPLE}.pair1.truncated -2 ${SAMPLE}.pair2.truncated -U ${SAMPLE}.collapsed,${SAMPLE}.collapsed.truncated,${SAMPLE}.singleton.truncated -S ${DIR}/${SAMPLE}.sam

cd ${DIR}

samtools view -bS ${SAMPLE}.sam | samtools sort > ${SAMPLE}.bam

rm ${SAMPLE}.sam
```

```
#!/bin/bash

#PBS -N 2020-06-12.bam_RG.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=48gb
#PBS -l walltime=12:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae
#PBS -J 01-49

cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/data/snp_variants


echo "working with sample:"
sed "${PBS_ARRAY_INDEX}q;d" /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/scripts/rawdata_processing/snp_variants/sampleorder_au.txt
SAMPLE=$(sed "${PBS_ARRAY_INDEX}q;d" /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/scripts/rawdata_processing/snp_variants/sampleorder_au.txt)

export _JAVA_OPTIONS="-Xmx48g"

java -Xmx48g -jar /apps/picard/2.18.26/picard.jar AddOrReplaceReadGroups INPUT=${SAMPLE}.bam OUTPUT=${SAMPLE}RG.bam RGID=1 RGLB=lib1 RGPL=illumina RGPU=${SAMPLE} RGSM=${SAMPLE}
```

```

#!/bin/bash

#PBS -N 2020-06-12.bam_duplicated.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=120gb
#PBS -l walltime=12:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae
#PBS -J 01-49

cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/data/snp_variants

OUT_DIR=/srv/scratch/canetoad/Stuart.Starling-Feb20/snp_variants

echo "working with sample:"
sed "${PBS_ARRAY_INDEX}q;d" /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/scripts/rawdata_processing/snp_variants/sampleorder_au.txt
SAMPLE=$(sed "${PBS_ARRAY_INDEX}q;d" /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/scripts/rawdata_processing/snp_variants/sampleorder_au.txt)

export _JAVA_OPTIONS="-Xmx120g"

java -Xmx48g -jar /apps/picard/2.18.26/picard.jar MarkDuplicates INPUT=${SAMPLE}RG.bam OUTPUT=${OUT_DIR}/${SAMPLE}_RGmark.bam METRICS_FILE=${OUT_DIR}/${SAMPLE}.metrics.txt MAX_FILE_HANDLES_FOR_READ_ENDS_MAP=1000;
```

since running HaplotypeCaller, don't need to realign or fix mates unless there is an error in ValidateSamFile 6 hrs

```

#!/bin/bash
```

```
#PBS -N 2021-06-14.bam_validate.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=48gb
#PBS -l walltime=12:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae
#PBS -J 01-49

cd /srv/scratch/canetoad/Stuart.Starling-Feb20/snp_variants

export _JAVA_OPTIONS="-Xmx48g"

echo "working with sample:"
sed "${PBS_ARRAY_INDEX}q;d" /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/scripts/rawdata_processing/snp_variants/sampleorder_au.txt
SAMPLE=$(sed "${PBS_ARRAY_INDEX}q;d" /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/scripts/rawdata_processing/snp_variants/sampleorder_au.txt)

java -Xmx48g -jar /apps/picard/2.18.26/picard.jar ValidateSamFile I=${SAMPLE}_RGmark.bam MODE=SUMMARY
```

grep "No errors found" 2021-06-14.bam_validate.pbs.o1238166.* | wc -l #everything looks redi 2 g0

index .bam files for HaplotypeCaller

```
#!/bin/bash

#PBS -N 2021-06-14.bam_index.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=80gb
#PBS -l walltime=12:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae
#PBS -J 01-49

cd /srv/scratch/canetoad/Stuart.Starling-Feb20/snp_variants

export _JAVA_OPTIONS="-Xmx80g"

echo "working with sample:"
sed "${PBS_ARRAY_INDEX}q;d" /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/scripts/rawdata_processing/snp_variants/sampleorder_au.txt
SAMPLE=$(sed "${PBS_ARRAY_INDEX}q;d" /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/scripts/rawdata_processing/snp_variants/sampleorder_au.txt)

java -Xmx80g -jar /apps/picard/2.18.26/picard.jar BuildBamIndex I=${SAMPLE}_RGmark.bam
```

prep the genome:

prepare the genome for GATK: index it (fai and dict files)
make dict file (sequence dictionary of the contig names)

```
cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/genome
module load samtools/1.10
module load java/8u121
module load picard/2.18.26
```

test if picard tools works

```
export _JAVA_OPTIONS="-Xmx10g"
java -Xmx10g -jar /apps/picard/2.18.26/picard.jar -h
```

```
java -Xmx10g -jar /apps/picard/2.18.26/picard.jar CreateSequenceDictionary R=Sturnus_vulgaris_2.3.1.simp.fasta
O=Sturnus_vulgaris_2.3.1.simp.dict
```

The above ran very fast (0.11 mins).

make fai index file to allow random access to fasta files

```
samtools faidx Sturnus_vulgaris_2.3.1.simp.fasta
```

likewise very shot <1 min run time

```
#!/bin/bash

#PBS -N 2021-06-15.bam_haplocaller.pbs
#PBS -l nodes=1:ppn=10
#PBS -l mem=120gb
#PBS -l walltime=48:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae
#PBS -J 01-49

module load samtools/1.10
module load java/8u121
module load gatk/4.1.0.0
module load picard/2.18.26

cd /srv/scratch/canetoad/Stuart.Starling-Feb20/snp_variants

export _JAVA_OPTIONS="-Xmx120g"

GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/genome/Sturnus_vulgaris_2.3.1.simp.fasta

echo "working with sample:"
sed "${PBS_ARRAY_INDEX}q;d" /srv/scratch/z5188231/KStuart.Starling-
Aug18/Sv5_AustraliaWGS/scripts/rawdata_processing/snp_variants/sampleorder_au.txt
SAMPLE=$(sed "${PBS_ARRAY_INDEX}q;d" /srv/scratch/z5188231/KStuart.Starling-
Aug18/Sv5_AustraliaWGS/scripts/rawdata_processing/snp_variants/sampleorder_au.txt)
```

```
gatk --java-options "-Xmx120G" HaplotypeCaller -R $GENOME -I ${SAMPLE}_RGmark.bam -O ${SAMPLE}.g.vcf -ERC GVCF
```

Then combine individual GVCF files into joint VCF

```
#!/bin/bash
```

```
#PBS -N 2021-06-15.variants_combine.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=120gb
#PBS -l walltime=100:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

module load samtools/1.10
module load java/8u121
module load gatk/4.1.0.0
module load picard/2.18.26

cd /srv/scratch/canetoad/Stuart.Starling-Feb20/snp_variants

export _JAVA_OPTIONS="-Xmx120g"

GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/genome/Sturnus_vulgaris_2.3.1.simp.fasta

gatk --java-options "-Xmx120G" CombineGVCFs \
--reference $GENOME \

--variant au01_lem.g.vcf \

--variant au02_lem.g.vcf \

--variant au03_mai.g.vcf \

--variant au04_mai.g.vcf \

--variant au05_men.g.vcf \

--variant au06_men.g.vcf \

--variant au07_won.g.vcf \

--variant au08_won.g.vcf \

--variant au09_mun.g.vcf \

--variant au10_hay.g.vcf \

--variant au11_mun.g.vcf \

--variant au12_con.g.vcf \

--variant au13_hay.g.vcf \

--variant au14_dub.g.vcf \

--variant au15_men.g.vcf \

--variant au16_mun.g.vcf \

--variant au17_mun.g.vcf \

--variant au18_con.g.vcf \

--variant au19_con.g.vcf \

--variant au20_hob.g.vcf \

--variant au21_hob.g.vcf \

--variant au22_hob.g.vcf \

--variant au23_dub.g.vcf \
--variant au24_dub.g.vcf \
--variant au25_hay.g.vcf \
--variant au26_lem.g.vcf \
--variant au27_mai.g.vcf \
--variant au28_mun.g.vcf \
--variant au29_mun.g.vcf \
--variant au30_won.g.vcf \
--variant au31_mun.g.vcf \
--variant au32_mun.g.vcf \
--variant au33_mun.g.vcf \
--variant us01_nyc.g.vcf \
--variant us02_nyc.g.vcf \
--variant us03_nyc.g.vcf \
--variant us04_nyc.g.vcf \
--variant us05_nyc.g.vcf \
--variant us06_nyc.g.vcf \
--variant us07_nyc.g.vcf \
--variant us08_nyc.g.vcf \
--variant uk01_nwc.g.vcf \
--variant uk02_nwc.g.vcf \
--variant uk03_nwc.g.vcf \
--variant uk04_nwc.g.vcf \
--variant uk05_nwc.g.vcf \
--variant uk06_nwc.g.vcf \
--variant uk07_nwc.g.vcf \
--variant uk08_nwc.g.vcf \
--output wgs_variantscombined.vcf
```

GenotypeGVCFS: https://gatk.broadinstitute.org/hc/en-us/articles/360037057852-GenotypeGVCFs

```
#!/bin/bash

#PBS -N 2021-06-28.variants_genotype.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=120gb
#PBS -l walltime=19:00:00
```

```
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

module load samtools/1.10
module load java/8u121
module load gatk/4.1.0.0
module load picard/2.18.26

cd /srv/scratch/canetoad/Stuart.Starling-Feb20/snp_variants
export _JAVA_OPTIONS="-Xmx120g"

GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/genome/Sturnus_vulgaris_2.3.1.simp.fasta

gatk --java-options "-Xmx120g" GenotypeGVCFs \
-R $GENOME \
-V wgs_variantscombined.vcf \
-O wgs_variantsgenotyped.vcf
```

Filtering GVCF files AUS

```
#!/bin/bash

#PBS -N 2021-07-02.variantfilter.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=56gb
#PBS -l walltime=12:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

GENOME=/srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/genome/Sturnus_vulgaris_2.3.1.simp.fasta

module load samtools/1.10
module load java/8u121
module load gatk/4.1.0.0
module load picard/2.18.26

cd /srv/scratch/canetoad/Stuart.Starling-Feb20/snp_variants

export _JAVA_OPTIONS="-Xmx56g"

gatk --java-options "-Xmx8g" VariantFiltration \
-R $GENOME \
-V wgs_variantsgenotyped.vcf \
-O wgs_variantsgenotyped_filtered.vcf \
--filter-name "first_snp_filter" \
--filter-expression "QD<2.0||FS>60.0||MQ<40.0||SOR>3.0"
```

Maf filter; depth filter; missing data filter (over total individuals)

--maf 0.1
--min-meanDP 2
--max-meanDP 50 #Avoid repetitive areas
--max-missing-count 4 #about 20% missing data

```
#!/bin/bash

#PBS -N 2021-07-02.variantfilter2.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=56gb
#PBS -l walltime=12:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

module load vcftools/0.1.16

cd /srv/scratch/canetoad/Stuart.Starling-Feb20/snp_variants

export _JAVA_OPTIONS="-Xmx56g"

vcftools --vcf wgs_variantsgenotyped_filtered.vcf --max-missing-count 4 --maf 0.1 --min-meanDP 2 --max-meanDP 50 --min-alleles 2 --max-alleles 2 --recode --out wgs_variantsgenotyped_filtered_maf01.vcf

vcftools --vcf wgs_variantsgenotyped_filtered.vcf --max-missing-count 4 --maf 0.05 --min-meanDP 2 --max-meanDP 50 --min-alleles 2 --max-alleles 2 --recode --out wgs_variantsgenotyped_filtered_maf005.vcf

vcftools --vcf wgs_variantsgenotyped_filtered.vcf --max-missing-count 4 --maf 0.01 --min-meanDP 2 --max-meanDP 50 --min-alleles 2 --max-alleles 2 --recode --out wgs_variantsgenotyped_filtered_maf001.vcf

vcftools --vcf wgs_variantsgenotyped_filtered.vcf --max-missing 0.5 --maf 0.03 --min-meanDP 2 --max-meanDP 50 --min-alleles 2 --max-alleles 2 --recode --out wgs_variantsgenotyped_filtered_miss50
```

SNPS: 19256335

LD Filter

https://www.biostars.org/p/338289/
pruned for LD by removing all sites with and r2 greater than 0.6 within 1kb sliding windows

```
#!/bin/bash

#PBS -N 2021-07-02.variantfilter3.pbs
#PBS -l nodes=1:ppn=16
#PBS -l mem=56gb
#PBS -l walltime=12:00:00
#PBS -j oe
#PBS -M katarina.stuart@student.unsw.edu.au
#PBS -m ae

module load samtools/1.9
module load vcftools/0.1.16

cd /srv/scratch/canetoad/Stuart.Starling-Feb20/snp_variants

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

bcftools +prune -l 0.6 -w 1000 wgs_variantsgenotyped_filtered_maf005.vcf.recode.vcf -Ov -o wgs_variantsgenotyped_filtered_maf005_r2.vcf
bcftools +prune -l 0.6 -w 1000 wgs_variantsgenotyped_filtered_miss50.recode.vcf -Ov -o  wgs_variantsgenotyped_filtered_miss50_r2.vcf
```

```
19256335


cd /srv/scratch/z5188231/KStuart.Starling-Aug18/Sv5_AustraliaWGS/data/snp_variants_processed/

grep "^#" -v wgs_variantsgenotyped_filtered_miss50.recode.vcf | wc -l

19256335

grep "^#" -v wgs_variantsgenotyped_filtered_miss50_r2.vcf | wc -l

7402600
```