**STEP 2:**Map to EquiCab 2.0 with Burrows-Wheeler Aligner

      — Remove ambiguously mapped reads, low quality reads and PCR duplicates

*Information from BWA website: BWA-MEM and BWA-SW share similar features such as long-read support and split alignment, but BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate. BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads.*

So try BWA-MEM

Download software: (into durwa004/Tools)

<https://sourceforge.net/projects/bio-bwa/files/bwakit/>

NB — is actually available (module load bwa/0.7.12)

# NB attempted to run without this and it kicked me out!

qsub -I -q ram1t -l nodes=1:ppn=8,mem=250gb,walltime=08:00:00

# Go to folder durwa004/PSSM2/2016\_PSSM2/WGS/

module load bwa/0.7.12

# Create reference sequence:

bwa index /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta

# Align sequence to reference:

# A2077:

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A2077\_S1\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A2077\_S1\_L001\_R2\_001.fastq > A2077.sam

# A2080

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A2080\_S3\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A2080\_S3\_L001\_R2\_001.fastq > A2080.sam

#A4229

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A4229\_S2\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A4229\_S2\_L001\_R2\_001.fastq > A4229.sam

#A4496

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A4496\_S4\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A4496\_S4\_L001\_R2\_001.fastq > A4496.sam

#A5576

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A5576\_S5\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A5576\_S5\_L001\_R2\_001.fastq > A5576.sam

#A5674

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A5674\_S11\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A5674\_S11\_L001\_R2\_001.fastq > A5674.sam

# A5675

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A5675\_S6\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A5675\_S6\_L001\_R2\_001.fastq > A5675.sam

# A5676

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A5676\_S8\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A5676\_S8\_L001\_R2\_001.fastq > A5676.sam

# A5677

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A5677\_S10\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A5677\_S10\_L001\_R2\_001.fastq > A5677.sam

# A5964

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A5964\_S7\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A5964\_S7\_L001\_R2\_001.fastq > A5964.sam

# A6099

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A6099\_S9\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121019\_M00784\_0021\_A000000000-A1L3A/A6099\_S9\_L001\_R2\_001.fastq > A6099.sam

#A2063

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A2063\_S2\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A2063\_S2\_L001\_R2\_001.fastq > A2063.sam

#A2068

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A2068\_S4\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A2068\_S4\_L001\_R2\_001.fastq > A2068.sam

#A2071

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A2071\_S5\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A2071\_S5\_L001\_R2\_001.fastq > A2071.sam

#A2072

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A2072\_S7\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A2072\_S7\_L001\_R2\_001.fastq > A2072.sam

#A2074

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A2074\_S9\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A2074\_S9\_L001\_R2\_001.fastq > A2074.sam

#A2129

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A2129\_S3\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A2129\_S3\_L001\_R2\_001.fastq > A2129.sam

#A3606

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A3606\_S8\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A3606\_S8\_L001\_R2\_001.fastq > A3606.sam

#A4033

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A4033\_S10\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A4033\_S10\_L001\_R2\_001.fastq > A4033.sam

#A4054

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A4054\_S11\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A4054\_S11\_L001\_R2\_001.fastq > A4054.sam

# A5659

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A5659\_S6\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A5659\_S6\_L001\_R2\_001.fastq > A5659.sam

# A598

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A598\_S1\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/miseq/121017\_M00784\_0020\_A000000000-A1LEB\_v2/A598\_S1\_L001\_R2\_001.fastq > A598.sam

# A2085

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/hiseq/121114\_SN261\_0464\_AC1CANACXX/Project\_McCue\_Project\_008/A2085\_GTGGCC\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/hiseq/121114\_SN261\_0464\_AC1CANACXX/Project\_McCue\_Project\_008/A2085\_GTGGCC\_L001\_R2\_001.fastq > A2085.sam

# A1543

bwa mem /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta /home/mccuem/data\_release/umgc/hiseq/121114\_SN261\_0464\_AC1CANACXX/Project\_McCue\_Project\_008/A1543\_GTTTCG\_L001\_R1\_001.fastq /home/mccuem/data\_release/umgc/hiseq/121114\_SN261\_0464\_AC1CANACXX/Project\_McCue\_Project\_008/A1543\_GTTTCG\_L001\_R2\_001.fastq > A1543.sam

Convert .sam to .bam files

# head Anumber.sam (if doesn't start with @):

**# samtools view -bT /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta Anumber.sam > Anumber.bam**

They do start with @ = continue with this:

module load samtools

head Anumber.sam (if does start with @):

**samtools view -bS Anumber.sam > Anumber.bam**

# Create shell script:

module load R/3.2.1

Rscript PSSM2\_shell\_script\_samtobam.R

# Run parallel using a pbs script: samtobam\_parallel.pbs

qsub -q ram256g samtobam\_parallel.pbs

# Then need to sort:

PSSM2\_Shell\_script\_bamtosorted.R

# Run shell script in parallel: bamtosorted\_parallel.pbs

qsub -q  ram256g bamtosorted\_parallel.pbs

For some reason 3606 didn't get converted to .bam so rerunning the .sam file then check head:

head A3606.sam # looks fine so do sam to bam:

samtools view -bS A3606.sam > A3606.bam

samtools sort A3606.bam A3606.sorted

# Need to index:

**samtools index FILENAME.sorted.bam** 

PSSM2\_shell\_script\_sortedtoindexed.R

SortedtoIndex.sh

qsub -q ram256g sortedtoindexed\_parallel.pbs

1916564.mesabim3.msi.umn.edu

# Index reference sequence:

**samtools faidx EquCab.fasta**

samtools faidx /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta

Calculate genotype likelihoods with mpileup and call variants using bcftools

**module load samtools**

**module load bcftools**

**samtools mpileup -ugf /home/mccuem/avilaf/FA/Arabians/EquCab.fasta AR\_COL140634.sorted.bam AR\_COL140627.sorted.bam AR\_COL140665.sorted.bam AR\_COL163140.sorted.bam AR\_COL140772.sorted.bam AR\_COL163145.sorted.bam Arabian2\_Be4.sorted.bam Arabian1\_Be1.sorted.bam AR\_COL140745.sorted.bam AR\_COL130059.sorted.bam AR\_COL130033.sorted.bam AR\_COL130030.sorted.bam Brooks\_2174.sorted.bam Arabian3\_GA50.sorted.bam | bcftools call -vmO v -o Arabians.vcf**

PSSM2\_shell\_script\_mpileup\_bcftools.R

mpileup\_bcftools.sh # Edited shell script so that it is in the above format (out = PSSM2\_QHs.vcf)

mpileup\_bcftools.pbs

1922964.mesabim3.msi.umn.edu # Error'd out — haven't checked why!!

# Needed reference ( /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn.fasta in the code (before the sequences)

**Error'd out:**

[mpileup] 24 samples in 24 input files

<mpileup> Set max per-file depth to 333

mpileup\_bcftools.pbs.e1931954 (END)

Added in -d 333 (after reference sequence)

1934839.mesabim3.msi.umn.edu

***Maybe here?***

Or, alternatively, in order to output a filtered VCF file:

**samtools mpileup -ugf /home/mccuem/avilaf/FA/Arabians/EquCab.fasta AR\_COL140634.sorted.bam AR\_COL140627.sorted.bam AR\_COL140665.sorted.bam AR\_COL163140.sorted.bam AR\_COL140772.sorted.bam AR\_COL163145.sorted.bam Arabian2\_Be4.sorted.bam Arabian1\_Be1.sorted.bam AR\_COL140745.sorted.bam AR\_COL130059.sorted.bam AR\_COL130033.sorted.bam AR\_COL130030.sorted.bam Brooks\_2174.sorted.bam Arabian3\_GA50.sorted.bam | bcftools view -bvcg - > Arabians.raw.bcf | bcftools view Arabians.raw.bcf | vcfutils.pl varFilter -D100 > Arabians.filtered.vcf**

VCF stats:

**bcftools stats Arabians.vcf > Arabians.vchk   
plot-vcfstats Arabians.vchk -p plots/**

GATK Haplotype Caller tutorial

You can find best practices for GATK's variant discovery pipeline can be found by [clicking here](https://www.broadinstitute.org/gatk/guide/bp_step.php?p=2). For advice on whether to call variants on your genomes together or individually, see their [FAQ page](https://www.broadinstitute.org/gatk/guide/article?id=4150).

Step 1: run Haplotype Caller on individual BAM files

GATK splits variant discovery into two separate steps: variant calling and variant filtration. I had a wide range of coverage depth (3-18X), so I split my variant calling into groups of samples with under 6X, 6-9X, and over 9X. GATK's Haplotype Caller is the first step of the variant calling process, and it works at the individual sample level. The following is the pbs script (+ comments) I used to carry out the variant calling step using haplotype caller on the samples with greater than 9X coverage:

#!/bin/bash -l  
# I asked for 16 processors because I was running this on Itasca where there is no node sharing.  
#PBS -l nodes=1:ppn=16,walltime=48:00:00  
cd ~/WGS/paleomix  
# riss\_util has lots of useful tools. The profile.pl tool tracks your CPU and memory usage for you.  
module load riss\_util  
profile.pl &  
module load gatk  
module load parallel  
module load java  
# The basic command for running Haplotype Caller on a single VCF is as follows:  
# java -jar GenomeAnalysisTK.jar -T HaplotypeCaller -R [FASTA file you mapped your reads to] -I [filename] --emitRefConfidence GVCF --variant\_index\_type LINEAR --variant\_index\_parameter 128000 -o [output filename].g.vcf  
# Here is the for loop piped to parallel command I used as an example of an option for parallelizing:  
for i in EGR1583.Equus\_cab\_nucl\_wChrUn1\_2.realigned.bam EGR356.Equus\_cab\_nucl\_wChrUn1\_2.realigned.bam EGR506.Equus\_cab\_nucl\_wChrUn1\_2.realigned.bam EGR510.Equus\_cab\_nucl\_wChrUn1\_2.realigned.bam EGR525.Equus\_cab\_nucl\_wChrUn1\_2.realigned.bam M1510.Equus\_cab\_nucl\_wChrUn1\_2.realigned.bam M1519.Equus\_cab\_nucl\_wChrUn1\_2.realigned.bam; do echo $i; done | parallel java -jar /home/mccuem/shared/bin/GenomeAnalysisTK.jar -T HaplotypeCaller -R /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn1\_2.fasta -I {} --emitRefConfidence GVCF --variant\_index\_type LINEAR --variant\_index\_parameter 128000 -o {}.g.vcf

Step 2: Genotype GVCFs

Once you have run Haplotype Caller on all of your samples individually, you will have a set of GVCFs as output. These are intermediate files that need to be fed into the part where you perform the actual genotyping. Here is the pbs script (+ comments) I used to do the genotyping step for the same samples shown above:

#!/bin/bash -l  
# Not sure if 150 hours is entirely necessary, but this was the longest queue on Itasca. Will double check and update this.  
#PBS -l nodes=1:ppn=1,walltime=150:00:00  
cd ~/WGS/paleomix  
module load riss\_util  
profile.pl &  
module load gatk  
module load parallel  
module load java  
# The basic command for running the Genotype GVCFs step is as follows:  
# java -jar GenomeAnalysisTK.jar -T GenotypeGVCFs -R [FASTA file you mapped your reads to] --variant [GVCF 1] --variant [GVCF 2] --variant [etc] -o [output filename].vcf  
java -jar /home/mccuem/shared/bin/GenomeAnalysisTK.jar -T GenotypeGVCFs -R /home/mccuem/shared/FASTAs/Equus\_cab\_nucl\_wChrUn1\_2.fasta --variant EGR1583.Equus\_cab\_nucl\_wChrUn1\_2.realigned.bam.g.vcf --variant EGR506.Equus\_cab\_nucl\_wChrUn1\_2.realigned.bam.g.vcf --variant EGR510.Equus\_cab\_nucl\_wChrUn1\_2.realigned.bam.g.vcf --variant EGR525.Equus\_cab\_nucl\_wChrUn1\_2.realigned.bam.g.vcf --variant M1510.Equus\_cab\_nucl\_wChrUn1\_2.realigned.bam.g.vcf --variant M1519.Equus\_cab\_nucl\_wChrUn1\_2.realigned.bam.g.vcf -o TB\_o9.vcf

Step 3: Recalibrate variant quality scores

The next major step in the process is where variant filtration happens. For this, we have to first recalibrate variant quality scores using a set of known variants. In our case, the best option is to compare against SNP chip data. We don't have a set of known indels like this, so I opted not to recalibrate and filter indels. Here is the pbs script I used to recalibrate variant quality scores after merging my three groups of genotyped samples: