SheepGenomesDB (SGD)

**Standardised Pipeline Guidelines**

**for Alignment and Quality Control of Raw Sequence Data and Variant Calling**

**Version 2.0**

Last modified: 13.10.2015

**Contacts:**

Rudiger Brauning [rudiger.brauning@agresearch.co.nz](mailto:rudiger.brauning@agresearch.co.nz)

Hans Daetwyler [hans.daetwyler@ecodev.vic.gov.au](mailto:hans.daetwyler@ecodev.vic.gov.au)

Sean McWilliam sean.mcwilliam@csiro.au

James Kijas [james.kijas@csiro.au](mailto:james.kijas@csiro.au)

# Contents

[0 Contents 1](#_Toc412721672)

[1 Scope 2](#_Toc412721673)

[2 Fastq file format 2](#_Toc412721674)

[3 Cleaning reads 2](#_Toc412721675)

[4 Choice of SGD ID 3](#_Toc412721676)

[5 Submitting clean sequence to SRA 4](#_Toc412721677)

[6 Choice of reference genome 4](#_Toc412721678)

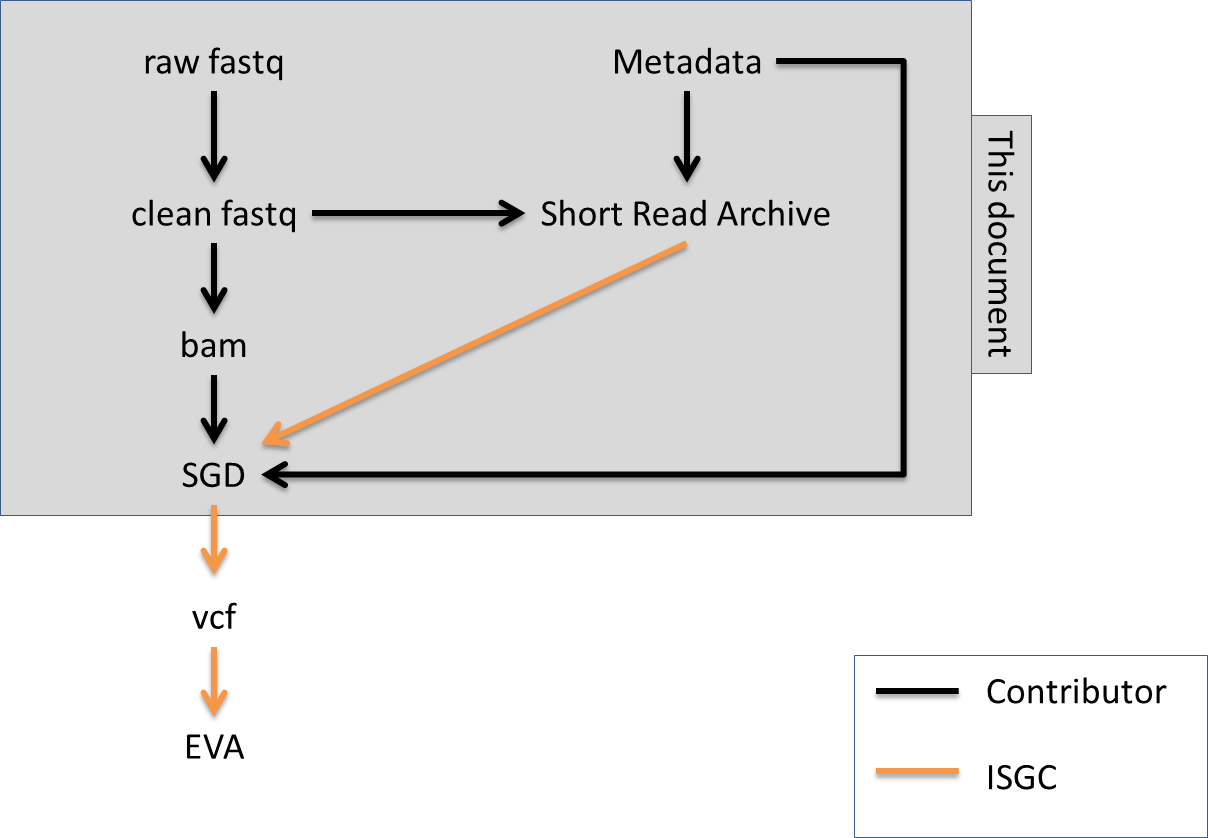
[7 Creation of alignment files (bam) 4](#_Toc412721679)

[8 SGD metadata 4](#_Toc412721680)

[9 References 5](#_Toc412721681)

# Scope

This document describes how to go from raw fastq files (sequence data) to cleaned bam files (sequence aligned against the reference genome). It also describes what additional information has to accompany a bam file in order for it to be considered for the SheepGenomesDB (SGD). The SGD is a project of the International Sheep Genomics Consortium. Periodically we will perform variant calling on all the aligned animals available. The resulting vcf files, together with metadata will then be submitted to the European Variation Archive (EVA) at Ensembl.



# Fastq file format

We require Illumina 1.8 (Phred+33) as fastq file format. For Illumina users, *CASAVA* version 1.8 produces fastq files with quality score as ASCII qscores Phred+33 encoded. If your data is from an older version of *CASAVA* and therefore Phred+64 encoded you can use the *–I* option when performing alignments with *bwa aln* to convert the ASCII qscores to Phred+33 encoding (this option is not available with *bwa mem*). Alternatively you can use *seqtk* to convert the qscores prior to alignment. For an overview of different fastq formats see <http://en.wikipedia.org/wiki/FASTQ_format>

Note: In the interest of quality control please do not merge fastq files across different lanes or flowcells.

# Cleaning reads

We highly recommend checking raw and filtered sequence reads using *FastQC* or equivalent.

The following sequential set of filters will reduce the number of false positive variants called from sequence pileups and must be performed on fastq files prior to alignment. We recommend running *quadtrim* with option *–d bulls*, which performs the above filtering steps on Illumina data. We also recommend *bio-stream-tools* which contain *nextseq-trim* for trimming strings of A and/or G from 3’ end of NextSeq reads. Further information can be requested from Amanda Chamberlain at [amanda.chamberlain@ecodev.vic.gov.au](mailto:amanda.chamberlain@ecodev.vic.gov.au).

1. Remove Illumina reads that do not pass Chastity filter. To check look at the read names, there should be an “N” present in the second field of the sequence ID line for “not filtered out” (@D00390:213:C4TY8ACXX:5:1101:2068:2113 1:N:0:4).
2. Remove adaptor sequence from reads.
3. Trim low quality bases (Phred score less than 20) from 5’ and 3’ ends.
4. Discard reads with a mean Phred score below 20.
5. Remove reads which have 3 or more N in the sequence.
6. Remove known artefacts. Reads from Illumina’s NextSeq machines are known to have strings of A and/or G 3’. These strings have normal qscores but are nothing more than artefacts.
7. Discard reads that are too short (less than 50% of their original length). Note: Where a read is left unpaired because its mate did not meet the required quality level, the orphaned read can still be aligned as a single ended read.

# Choice of SGD ID

To facilitate collaborative use of animal data while retaining some level of privacy we require a SGD ID to be chosen for each of your animals. This ID looks like the following:

<country of origin> <breed> <gender> <number>

A New Zealand texel ewe with number 12345 would have an SGD ID of:

NZ TXL F 000000012345

Build your SGD ID step by step:

1. For two letter country codes see <http://en.wikipedia.org/wiki/ISO_3166-1_alpha-2>
2. Breed abbreviations can be found on <http://sheepgenomesdb.org/>   
   There are several scenarios:
3. Main breed component (e.g. ROM) >=70%

Suggestion: <ROM>, record other breed components as metadata

1. Main breed component <50% (e.g. 50% Romney, 25% East Friesian, 25% Texel)

Suggestion: <CMP>, record other breed components as metadata. <CMP> stands for ‘composite’.

1. Composites with special names, e.g. Primera, Highlander

Suggestion: <CMP>, record special name and breed components as metadata

1. Breed unknown

Suggestion: <UNK>

1. For gender choose between F and M.
2. Finally choose a number (this could e.g. be a national ID) to complete the SGD ID. Zero fill this number to make up 12 digits.
3. To ensure uniqueness of SGD IDs please send your desired SGD IDs to one of the contacts listed above.

# Submitting clean sequence to SRA

We recommend submitting your clean fastq sequences to NCBI’s Sequence Read Archive ([SRA](http://www.ncbi.nlm.nih.gov/sra/)).

# Choice of reference genome

Oar\_v3.1 from Ensembl is the strongly preferred reference genome file. Standardising the reference file across all collaborators ensures minimal issues with combining the data. The reference genome can be downloaded from [ftp://ftp.ensembl.org/pub/release-78/fasta/ovis\_aries/dna/Ovis\_aries.Oar\_v3.1.dna\_sm.toplevel.fa.gz](ftp://ftp.ensembl.org/pub/release-78/fasta/ovis_aries/dna/Ovis_aries.Oar_v3.1.dna_sm.toplevel.fa.gz%20)

The reference genome contains the 26 autosomes, the X chromosome, the unassigned scaffolds, and the mitochondrial genome. We encourage you to compare the checksum of the downloaded file with the checksum provided at Ensembl (<ftp://ftp.ensembl.org/pub/release-78/fasta/ovis_aries/dna/CHECKSUMS>). Hint: use the UNIX *sum* command.

Note: While the sequence data for Oar\_v3.1 is identical across Ensembl, NCBI, and UCSC, the contig names are not. Each site requires their particular naming scheme for things to work with their genome browsers or variant effect predictors.

# Creation of alignment files (bam)

Please follow the steps below to get to a single bam file per animal:

1. Choose correct reference genome (described above).
2. Align fastq files to reference using *bwa mem* 0.5.9+ with default parameters. *bwa mem* is faster and more accurate than *bwa aln*. Reads have to be at least 70bp in length for *bwa mem*.
3. Within *bwa mem* use *–M* to ensure *Picard* compatibility.
4. Within *bwa* use *–R* to set RG headers, ensuring that sample (SM) field contains the SGD ID.
5. Do not filter alignments by mapping quality. We will do that when we call variants.
6. Merge bam files where they correspond to the same individual. This can be done with *MergeSamFiles* (*picard*). Ensure the SGD ID is in the final bam file name.Remove or mark PCR duplicates. This can be done using *rmdup* (*samtools*) or *MarkDuplicates* (*picard*).
7. Perform a local realignment around indels to minimise mismatches across all reads. This can be done with the *GATK* tools *RealignerTargetCreator* and *IndelRealigner*.
8. If you haven’t performed step 4 (specifying SM field in RG header lines as SGD ID), fix this now using *samtools* *reheader*.
9. Sort alignments in order of correct reference genome (described above).
10. Create index files (.bai).
11. Calculate the final mean coverage of the alignment. This can be done using *DepthOfCoverage* (*GATK*). Include coverage statistics in the metadata list (described below).

# SGD metadata

Metadata provides information on the animals and is essential to the use of the SGD. The minimum information required is:

* Breed country
* Breed
* Gender
* SGD ID
* Contributor (name, institution, address, email)

The recommended additional information is:

* Animal
  + NCBI’s Biosamples ID (if available)
  + Tissue, e.g. blood
  + National ID (if available)
  + Aliases
  + Age
  + Year of birth
  + Country
  + GPS location coordinates
  + Health
  + Picture
* Sequencing
  + Platform, e.g. Illumina HiSeq 2500
  + Version, e.g. v4 chemistry
  + SRA accession number (if available)
  + DNA extraction method (high salt, filter, magnetic beads)
  + Mean coverage depth
* Analysis
  + Software, software version, commands
* Genotypes
  + Pointer to other genotype information for this animal, e.g. HD SNP chip results
* Funding
  + Keep your funding agency happy and mention them here

It’s not a problem to include additional categories, in fact this is encouraged.

# Variant Calling

Run 1 of SGD will use both Samtools1.2 and the GATK Unified Genotyper to independently call variants. The intersect set will be defined as the collection of ‘unfiltered’ variants transfer to EBI. Variants will then be filtered to remove low quality loci, and a second collection of ‘filtered’ variants made available to EBI.

Only filtered variant sets will be interfaced with dbSNP. Filtered variants will be searchable via programs on SGD website SheepGenomesDB.org that connect to EVA.

## In pipeline order:

**Samtools command line:**

samtools-1.2/samtools mpileup -r Chr1:0-20000000 -P ILLUMINA,illumina,Illumina,SOLiD –BA -ugf ReferenceFile -b BamListFile –t DP,DPR,DV,DP4,INFO/DPR,SP | bcfTools call –mv -> out.vcf

**GATK UG command line:**

java -Xmx90g -jar GenomeAnalysisTK.jar -R ReferenceFile -T UnifiedGenotyper BamFileList -o OutPutFile -out\_mode EMIT\_VARIANTS\_ONLY -stand\_call\_conf 20 -stand\_emit\_conf 20 -A FisherStrand -A StrandOddsRatio -A StrandBiasBySample -rf BadCigar --genotype\_likelihoods\_model BOTH -L Chr1:0-20000000 –log UG.log

*All the following filters implemented in python using open source python parser PyVCF* [*https://github.com/jamescasbon/PyVCF/*](https://github.com/jamescasbon/PyVCF/) *to read vcf file. Downloaded 29.03.2015.*

## Removal of variants with 2 or more alternative alleles

* Unfiltered vcf file of variants with 2 or more alternative alleles available upon request

## Minimum number of alternative allele observations on forward and reverse reads

* Threshold used was 1, removes variants never observed on forward or reverse reads
* Replaces strand bias filters, which was found to be too dependent on sample size making it cumbersome to choose an optimal p-value

## Overall quality

* samtools mpileup populates the vcf file QUAL field.
* Threshold chosen QUAL 20 (phred score)

## Mapping quality

* Filtered on MQ 30 (phred)

## Setting minimum and maximum read depths for filtering

* Set minimum number as 10 across all animals
* Set maximum as: median read depth + 3 \* standard deviation read depth

## Opposing Homozygotes Filter

* Count the number of opposing homozygotes in parent/offspring pairs
  + Accumulate information as proportion of homozygotes per SNP and individual
* Use a threshold to filter variants that produce too many opposing homozygotes
  + Used 10%
* Implemented in bash and Fortran90

## Remove variants with same basepair position

* If two variants have the same bp position both are removed
* Resolves issue with SNP and INDEL calls at same position

## Filtering INDELs based on proximity to INDELs

* Remove lower QUAL INDEL when closer than 10 basepairs

## Filtering all variants based on proximity to each other

* remove lower QUAL variants if closer than 3bp

## Filtering SNP based on proximity to INDELs

* Remove SNP closer than 5bp to INDELs

## Creating union sets

A union set of the Samtools and GATK called variants was produced using GATK CombineVariants to produce the final VCF files. This was conducted independently for unfiltered and filtered variant calls. An intersect set, for the common variant locations predicted by both Samtools and GATK, was produced using GATK SelectVariants on the filtered union set. This filtered intersect set is considered as the high quality variants for submission to EVA.

***Union set command line:***

java -jar /apps/gatk/3.5.0/GenomeAnalysisTK.jar -T CombineVariants -R $ref --variant:samtools $vcf1 --variant:gatk $vcf2 -o gatk\_merged.${chr}.vcf -genotypeMergeOptions PRIORITIZE -priority gatk,samtools

***Intersect set command line:***

java -jar /apps/gatk/3.4.46/GenomeAnalysisTK.jar -T SelectVariants -R $ref -V:variant gatk\_merged.${chr}.vcf -select 'set == "Intersection";' -o gatk\_intersect.${chr}.vcf

# References

* bio-stream-tools  
  <https://bitbucket.org/arobinson/bio-stream-tools/overview>
* bwa  
  Li H and Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.
* FastQC  
  http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
* GATK  
  Depristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43, 491–498. doi:10.1038/ng.806
* picard  
  <http://picard.sourceforge.net/index.shtml>
* quadtrim  
  <https://bitbucket.org/arobinson/quadtrim>
* samtools  
  Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9.
* PyVCF – Python VCF file parser

<https://github.com/jamescasbon/PyVCF/>

* Samtools1.2

<http://www.htslib.org/doc/samtools.html>

* seqtk  
  <https://github.com/lh3/seqtk>