Sequence Alignment Guidelines for producing bam files for the 1000 bull genomes project

Version: 19.12.2013

ASCII qscores in bam files are required to be Phred+33 encoding.(Required)

For Illumina users, CASSAVA 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 using the –I option when performing alignments in BWA will convert the ASCII qscores to Phred+33 encoding.

Filtering of reads before alignment (Required)

The following filters will reduce the number of false positive variants called from sequence pileups and must be performed on fastq files.

1. Remove reads that do not pass Chastity (for Illumina sequence data).
2. Remove reads which have 3 or more N in the sequence
3. Remove reads with and a mean Phred quality score of less than 20.
4. Adaptor sequence must be trimmed from reads.
5. Low quality bases (with Phred score less than 20) must then trimmed from the 5' and 3` end,
    a. If after the trimming the read does not meet the above quality standard (average Phred score 20) it must be discarded.
    b. If after trimming a read is less than 50% of standard length for technology it must be discarded (i.e. standard length for Illumina is 100 so minimum should be 50 bases)

Note: Where a read is left unpaired because its mate did not meet the required quality level, the read can be aligned as a single ended read.

We highly recommend partners check raw and filtered sequence reads in a program FastQC (or equivalent) which can be downloaded from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ . We also recommend a script that performs the above filtering steps on Illumina data which can be downloaded from https://bitbucket.org/arobinson/qualitytrim . Further information can be requested from Amanda Chamberlain: amanda.chamberlain@depi.vic.gov.au.

Reference Sequence for alignment (Required)

UMD3.1 is the reference genome to be used in this project. The correct copy of the reference genome must be used to ensure your bam files are compatible with the analysis pipeline. It can be downloaded from project partner Paul Stothard here: http://stothard.afns.ualberta.ca/1000_bull_genomes/reference_for_mapping/umd_3_1_reference_1000_bull_genomes.fa.gz or via link on http://www.1000bullgenomes.com

Alignment and bam file creation(Required)

We used BWA version 0.5.9 for mapping reads to the reference genome. For the alignment the default BWA parameters are used. SAMTOOLS can then be used to create bam files. The following must be done during this process:

1. PCR duplicates removed or marked. This can be done using rmdup tool of samtools or MarkDuplicates tool of picard.
2. Aligned to correct reference genome (described above).
3. Merge bam files where they correspond to the same individual, this can be done with Picard's MergeSamFiles.jar, and ensure the international ID is in the final bam file name.
4. bam files must be sorted (in order Chr1, Chr2, Chr3, …, Chr29, X, other contigs) and indexed, with the index file (.bai) included on the hard drive.
5. RG header lines must contain the international ID of the animal in the sample (SM) field. This can be done using reheader tool in samtools.
6. Calculate the final mean coverage of the alignment. This can be done using the DepthOfCoverage tool in the Genome Analysis Toolkit (GATK). Include coverage statistics in the checklist (described below).

Optional modifications

We suggest a local realignment around indels to minimise mismatches across all reads. This can be done with GATK tools RealignerTargetCreator and IndelRealigner. Note: this is optional

Optional additional data

The inclusion of genotype information aids the quality checking process and can identify problems with libraries or alignments. If you have BovineSNP50 or BovineHD data for your samples it would be very beneficial to include these. Genotype data file should be the Illumina GenomeStudio output in TOPTOP (preferred) and FORWARD/FORWARD format. Please contact us if you have Affymetrix genotype data.

Sending bam files(Required)

Md5 sum files must be created for all files sent on disk drive. This allows us to check whether data has transferred correctly. Sorted bam files, their associated indexes, md5 sums files and genotype files (optional) must be copied to an external USB disk drive (USB3 and without power plug preferred). Please label the drive with your name and institution and then send to:

Dr. Hans Daetwyler
AgriBio Centre
Department of Environment and Primary Industries
5 Ring Rd.
Bundoora 3083
Victoria, Australia

1000 bull genomes checklist(Required)

Consult the 1000 bull genomes file submission checklist (available at http://www.1000bullgenomes.com) **before** preparing data. Checklist must then be filled in and submitted via email to Amanda Chamberlain, Hans Daetwyler and Ben Hayes once drives are in the mail (amanda.chamberlain@depi.vic.gov.au , ben.hayes@depi.vic.gov.au , hans.daetwyler@depi.vic.gov.au).

NOTE: bam files will not be accepted if they do not meet theses specifications.

We realise this is all computationally demanding, and so we can do some or all steps for you if required. Please contact us to discuss what help is available.

International Bull IDs of those bulls already included in the project are added into the Project Database hosted by project partner Bernt Guldbrandtsen (email for access Bernt.Guldbrandtsen@agrsci.dk).