

MB-GATK-SGE pipeline

For GATK best practices: classic UG / v3.x HC / MuTect
Matthew Bashton

Classic Unified Genotyper workflow

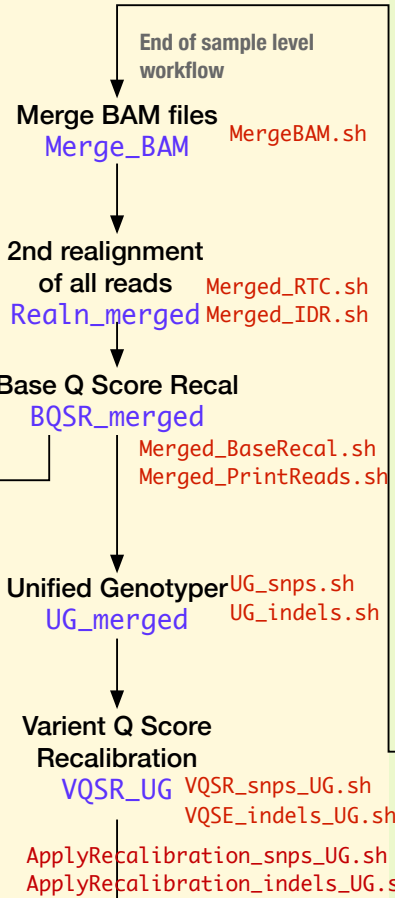
BAM files merged using Picard threading used to off-load (de)compression/IO, shell script takes path/*.*.bam as input from command line

Reads are realigned around indels, two stages:
i) Realignment Target Creation,
ii) Indel Realignment

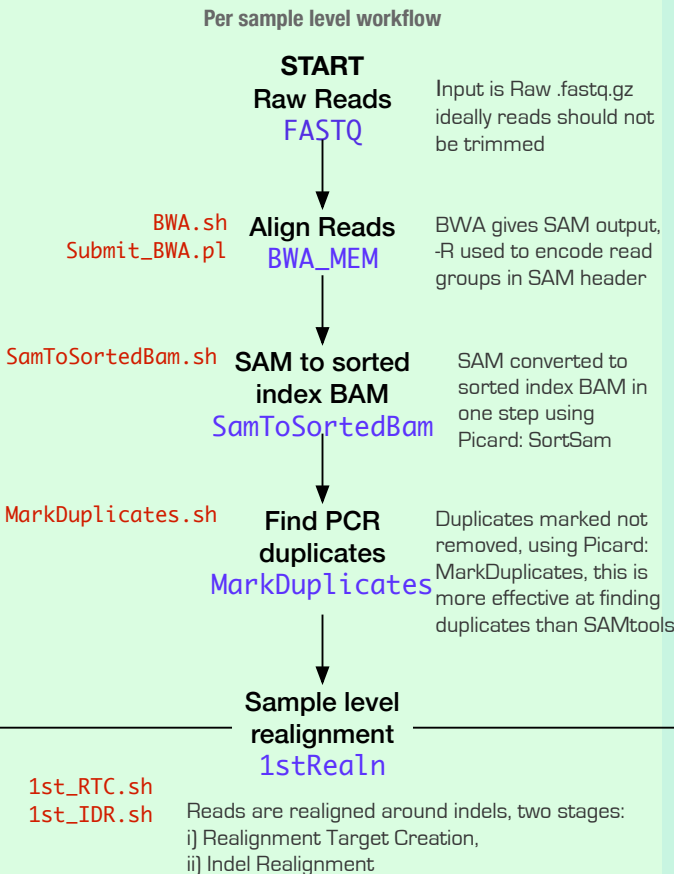
Q scores for each base are recalibrated using machine learning. Two stages i) build model ii) apply it and "print" a new set of reads

Variants called on all samples simultaneously, using Unified Genotyper, calls SNPs and indels separately owing to size of unified dataset

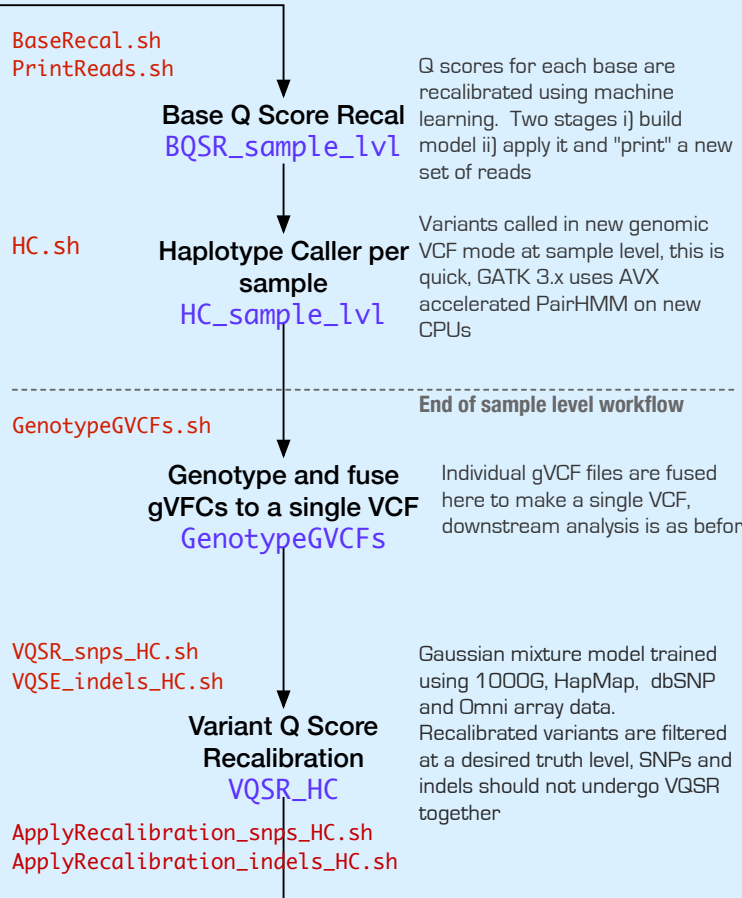
Gaussian mixture model trained using 1000G, HapMap, dbSNP and Omni array data. Recalibrated variants are filtered at a desired truth level, SNPs and indels should not undergo VQSR together



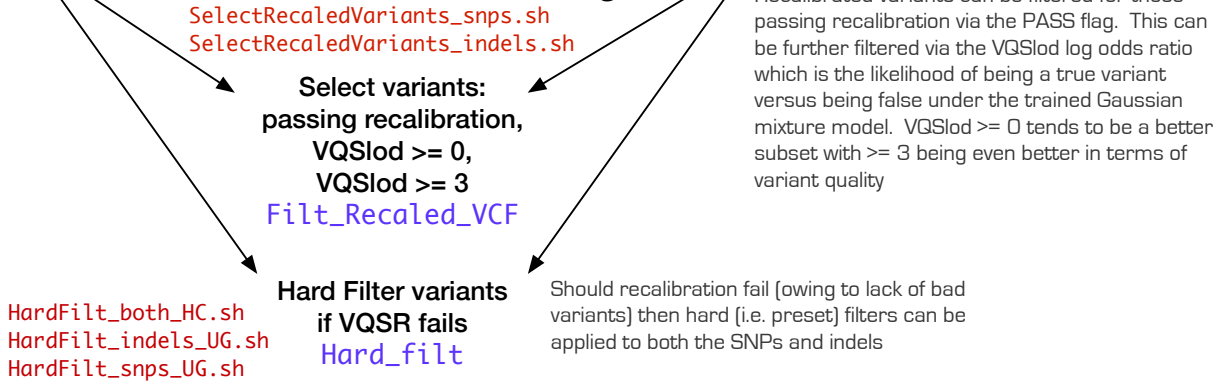
Common per-sample processing



New Haplotype Caller workflow



Recalibrated variant filtering



MuTect 1.x somatic variant calling

MT.sh
Submit_MT.pl
Call tumor / normal pairs using MuTect
MuTect

MuTect subtracts the normal (germline) variants from the tumor (somatic) variants. It also reports if SNPs are novel i.e. not in COSMIC or dbSNP

Perl script submits MuTect jobs from a list of paired normal/tumour BAM files

MuTect handles heterogeneous and impure tumour samples