

MB-GATK-SGE pipeline

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

Classic Unified Genotyper workflow

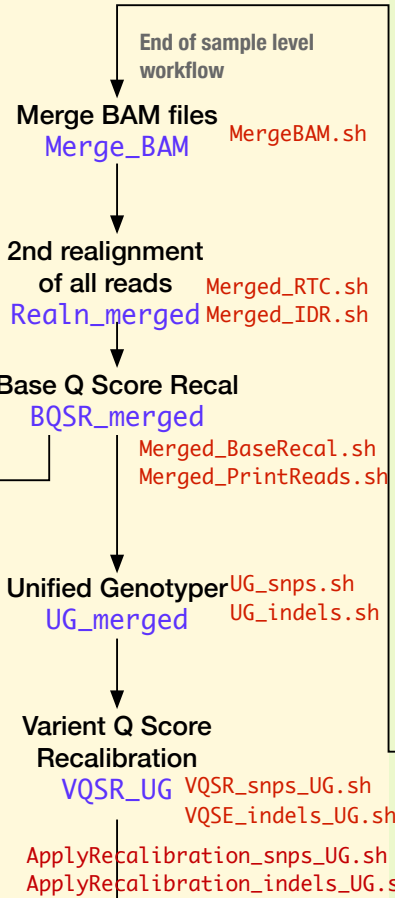
BAM files merged using Picard threading used to off-load (de)compression/IO, shell script takes path/*.bam as input from command line

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

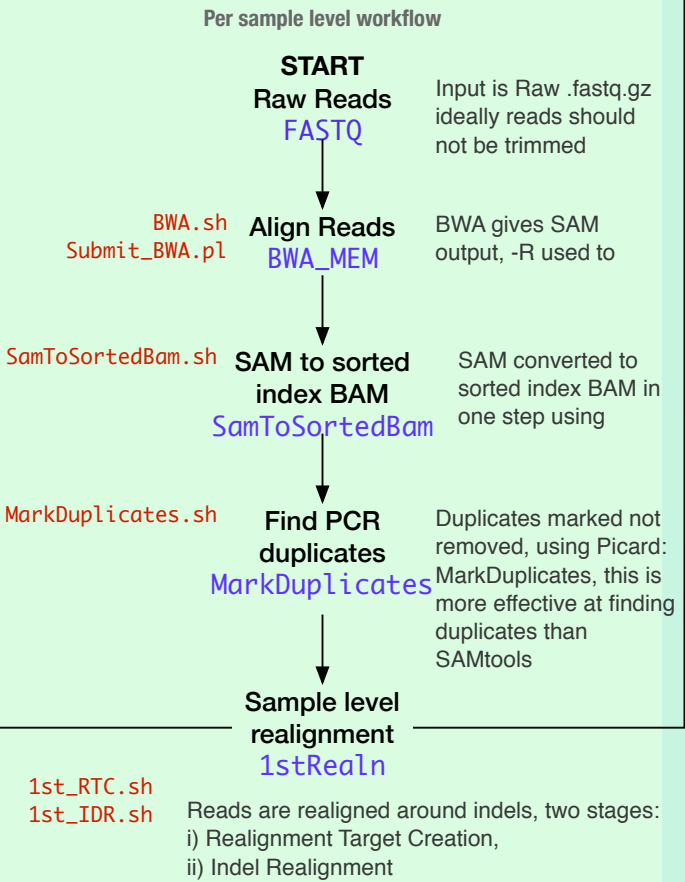
Q scores for each base are recalibrated using machine learning. Two stages i) build model ii) apply it and "print" a

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

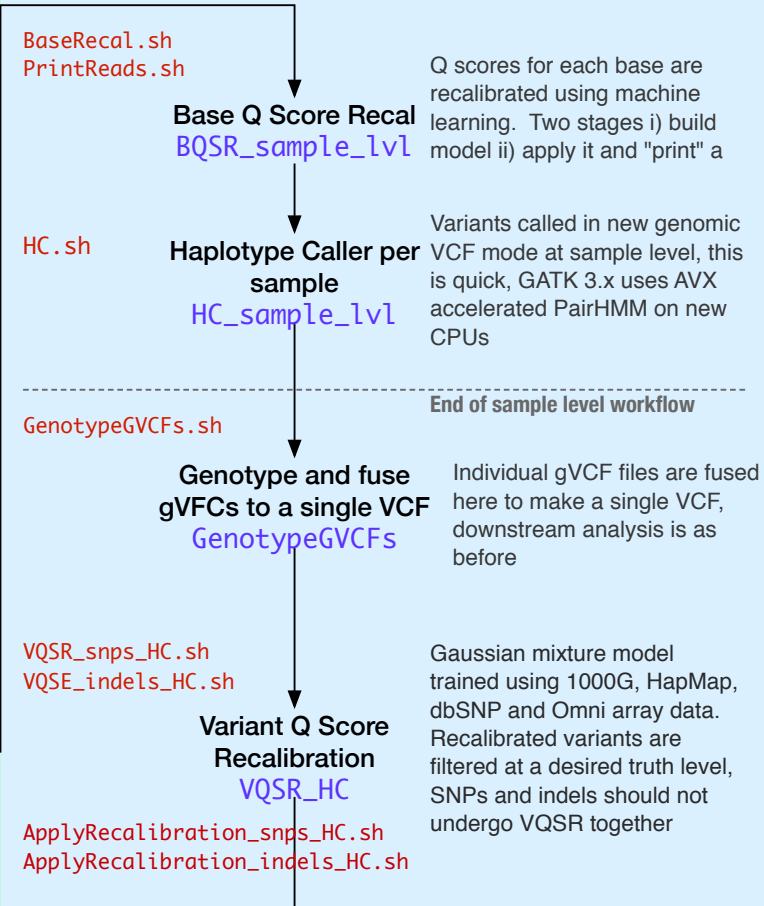
Gaussian mixture model trained using 1000G, HapMap, dbSNP and Omni array data. Recalibrated variants are filtered at a desired truth level, SNPs and



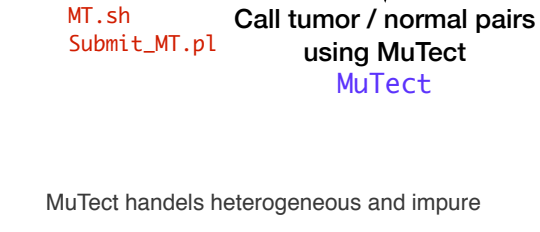
Common per-sample processing



New Haplotype Caller workflow



MuTect 1.x somatic variant calling

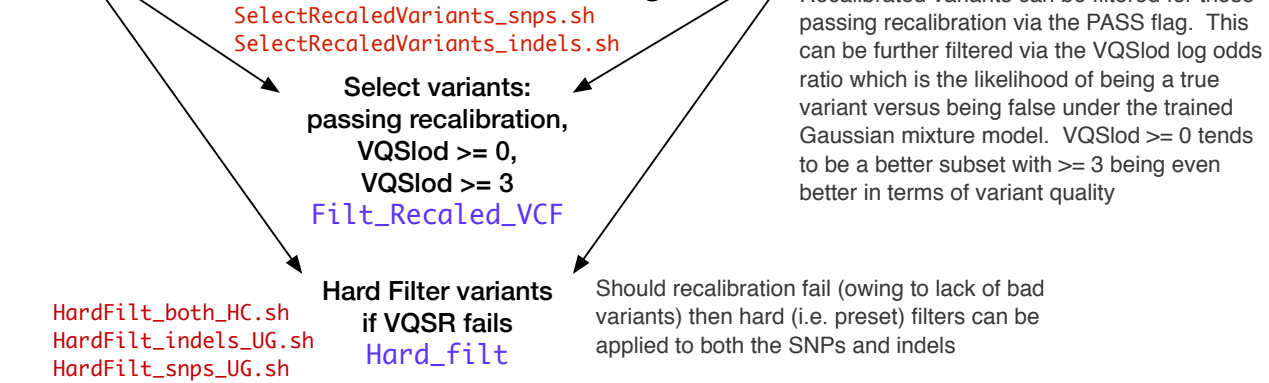


MuTect subtracts the normal (germline) variants from the tumor (somatic) variants. It also reports if SNPs are novel i.e. not in COSMIC or dbSNP

MuTect handles heterogeneous and impure

Perl script submits MuTect

Recalibrated variant filtering



Recalibrated variants can be filtered for those passing recalibration via the PASS flag. This can be further filtered via the VQSlod log odds ratio which is the likelihood of being a true variant versus being false under the trained Gaussian mixture model. VQSlod >= 0 tends to be a better subset with >= 3 being even better in terms of variant quality

Should recalibration fail (owing to lack of bad variants) then hard (i.e. preset) filters can be applied to both the SNPs and indels