MB-GATK-SGE pipeline

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

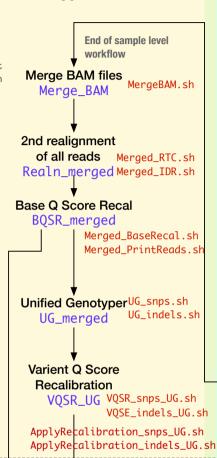
Reads are realigned around indels, two stages:

- i) Realignment Target Creation,
- ii) Indel Realianment

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 Per sample level workflow START Input is Raw .fasta.az Raw Reads ideally reads should not FASTO. be trimmed Align Reads BWA gives SAM output, Submit_BWA.pl -R used to encode read **BWA MFM** groups in SAM header SamToSortedBam.sh SAM to sorted SAM converted to sorted index BAM in index BAM one step using SamToSortedBam Picard: SortSam MarkDuplicates.sh Find PCR Duplicates marked not removed, using Picard: duplicates MarkDuplicates MarkDuplicates, this is more effective at finding duplicates than SAMtools Sample level realignment 1stRealn RTC.sh Reads are realigned around indels, two stages: IDR.sh i) Realignment Target Creation, ii) Indel Realianment

HardFilt_both_HC.sh

HardFilt_snps_UG.sh

HardFilt_indels_UG.sh

New Haplotype Caller workflow BaseRecal.sh Q scores for each base are PrintReads.sh recalibrated using machine Base Q Score Recal learning. Two stages i) build BOSR_sample_lvl model ii) apply it and "print" a new set of reads Variants called in new genomic HC.sh Haplotype Caller per VCF mode at sample level, this is quick. GATK 3.x uses AVX sample accelerated PairHMM on new HC_sample_lvl CPI Is End of sample level workflow GenotypeGVCFs.sh Genotype and fuse Individual qVCF files are fused here to make a single VCF, gVFCs to a single VCF downstream analysis is as before GenotypeGVCFs VOSR_snps_HC.sh Gaussian mixture model trained VOSE indels HC.sh using 1000G, HapMap, dbSNP and Omni array data. Variant Q Score Recalibrated variants are filtered Recalibration at a desired truth level. SNPs and indels should not undergo VQSR VOSR HC together ApplyRecalibration_snps_HC.sh ApplyRecalibration_indels_HC.sh

Recalibrated variants can be filtered for those

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

subset with >= 3 being even better in terms of

mixture model. VQSlod >= 0 tends to be a better

passing recalibration via the PASS flag. This can

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 iobs from a list of paired normal/tumour BAM files

SelectRecaledVariants snps.sh SelectRecaledVariants_indels.sh Select variants: passing recalibration, VQSlod >= 0, VQSlod >= 3Filt_Recaled_VCF

Hard Filter variants

if VQSR fails

Hard filt

Recalibrated variant filtering

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

variant quality

MuTect handels heterogeneous and impure tumour samples