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 Realignment

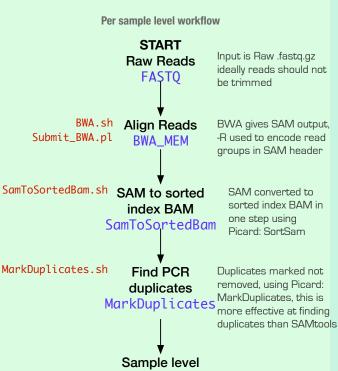
Q scores for each base are recalibrated using machine learning. Two stages i) build model ii) apply it and "print" a new set of reads

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

End of sample level workflow Merge BAM files Merae_BAM 2nd realignment of all reads Merged_RTC.sh Realn_merged Merged_IDR.sh Base Q Score Recal BOSR_merged Merged_BaseRecal.sh Merged_PrintReads.sh Unified Genotyper UG_snps.sh UG_indels.sh UG_merged Varient Q Score Recalibration VOSR UG VQSR_snps_UG.sh VQSE_indels_UG.sh ApplyRecalibration_snps_UG.sh ApplyRecalibration_indels_UG.sh

Common per-sample processing



realianment

1stRealn

i) Realignment Target Creation,

ii) Indel Realignment

Reads are realigned around indels, two stages:

New Haplotype Caller workflow

PrintReads.sh Q scores for each base are 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 auick. GATK 3.x uses AVX sample accelerated PairHMM on new HC_sample_lvl **CPUs**

GenotypeGVCFs.sh Genotype and fuse Individual gVCF files are fused

gVFCs to a single VCF

GenotypeGVFCs

VOSR_snps_HC.sh VOSE_indels_HC.sh Variant Q Score Recalibration

BaseRecal.sh

VOSR HC ApplyRecalibration_snps_HC.sh ApplyRecalibration_indels_HC.sh Gaussian mixture model trained using 1000G, HapMap, dbSNP and Omni array data.

End of sample level workflow

here to make a single VCF,

downstream analysis is as before

Recalibrated variants are filtered at a desired truth level. SNPs and indels should not undergo VQSR together

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

1st_RTC.sh

1st_IDR.sh

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

Recalibrated variant filtering SelectRecaledVariants snps.sh

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

Hard_filt HardFilt_snps_UG.sh

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

Hard Filter variants Should recalibration fail (owing to lack of bad HardFilt_both_HC.sh variants) then hard (i.e. preset) filters can be if VQSR fails HardFilt_indels_UG.sh applied to both the SNPs and indels

MuTect handels heterogeneous and impure tumour samples