## **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

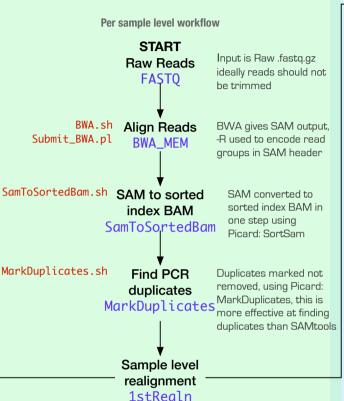
Q scores for each base are recalibrated using machine learning. Two stages i] build model ii] apply it and "print" a new set of reads

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\_IDR.sh Base Q Score Recal BQSR\_merged Merged\_BaseRecal.sh Meraed 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



Reads are realigned around indels, two stages:

i) Realignment Target Creation,

HardFilt\_snps\_UG.sh

ii) Indel Realignment

#### **New Haplotype Caller workflow**

BaseRecal.sh
PrintReads.sh

Base Q Score Recal

BQSR\_sample\_lvl

BQSR\_sample\_lvl

BQSR\_sample\_lvl

BQSR\_sample\_lvl

BQSR\_sample\_lvl

BQSR\_sample\_lvl

Variants called in new genomic

Haplotype Caller per VCF mode at sample level, this is quick, GATK 3.x uses AVX accelerated PairHMM on new CPUs

GenotypeGVCFs.sh

Genotype and fuse gVFCs to a single VCF
GenotypeGVCFs

Individual gVCF files are fused here to make a single VCF, downstream analysis is as before

End of sample level workflow

VQSR\_snps\_HC.sh VQSE\_indels\_HC.sh

HC.sh

Variant Q Score Recalibration VOSR\_HC

ApplyRecalibration\_snps\_HC.sh
ApplyRecalibration\_indels\_HC.sh

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

indels show

## 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

Select variants:

passing recalibration, VQSlod >= 0, VQSlod >= 3 Filt\_Recaled\_VCF

HardFilt\_both\_HC.sh
HardFilt\_indels\_UG.sh
HardFilt\_appa NC.sh
Hard\_filt

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

Recalibrated variants can be filtered for those

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

MuTect handels heterogeneous and impure tumour samples