Genomics Data Analysis

Professor Seunggeun Lee

Spring 2021

GATK best practice session

TA Jangho Kim

Table of Contents

1. Prerequisite (software)

- Computation with Linux OS (GSDS cluster)
- Conda virtual environment recommended
  - java
  - picard with java https://broadinstitute.github.io/picard/
  - samtools
  - gatk https://github.com/broadinstitute/gatk/releases
  - Burrows Wheeler Aligner https://sourceforge.net/projects/bio-bwa/

1. Prerequisite (files)

- fasta file of reference genome (GRCh 37)

  human_g1k_v37.fasta

- VCF of known variations

  ALL.wgs.mergedSV.v8.20130502.svs.genotypes.vcf.gz

  ALL.wgs.mergedSV.v8.20130502.svs.genotypes.vcf.gz.tbi

Downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/

Sample Code

```
conda activate [r_env]    # your environment name

conda install -c anaconda openjdk

conda install -c bioconda samtools

export PATH=$PATH:/home/smkjh1028/gatk-4.2.0.0

export PATH=$PATH:/home/smkjh1028/bwa-0.7.17

samtools faidx human_g1k_v37.fasta

gatk CreateSequenceDictionary -R human_g1k_v37.fasta

bwa index -a bwtsw human_g1k_v37.fasta
```

FASTQ

```
@SRR062641.1
GAAAGAAAGAAAGTCAACTGTATGCTTAAAAATCCAAGTTGTGGGTGGGAAGCTGAT
TGAATTTTTTACTACGGTTCATAAAAAAACACAAGACTCACAT
+
@7:6+)=0577''3=?->>;:A>A##################################
########################################
@SRR062641.10
AAGGGAACTGAAGTTTCTTCCATGTTGTTTACAGTATTACTTCAAAAAGTAAGTCCT
TTGGCATGTTCAGGTGCGACAACACCAGAGGAAGAATCTTACA
+
CAAC?BD?B5D5@=BDD=@D5C:DB?D?AA);>@########################
########################################
```

SAM/BAM file

HEADER lines starting with @ symbol describing various metadata for *all* reads

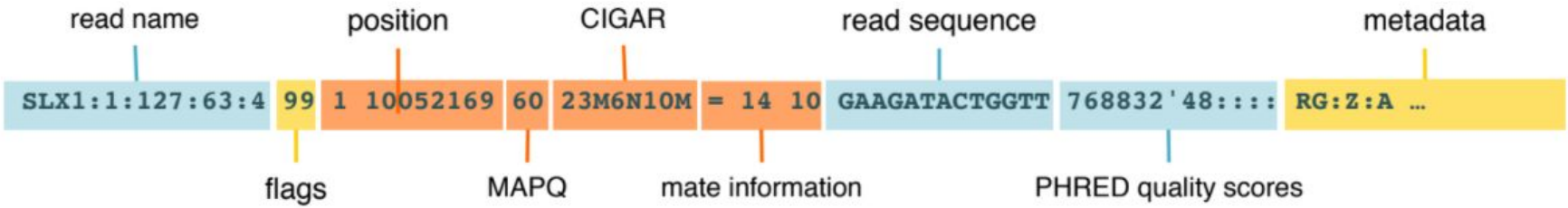
```
@HD VN:1.6 SO:coordinate
@SQ SN:seq1 LN:394893
@SQ SN:seq2 LN:92783
@RG ID:A SM:SAMPLE_A
```

BAM header line

Reference sequence dictionary entries

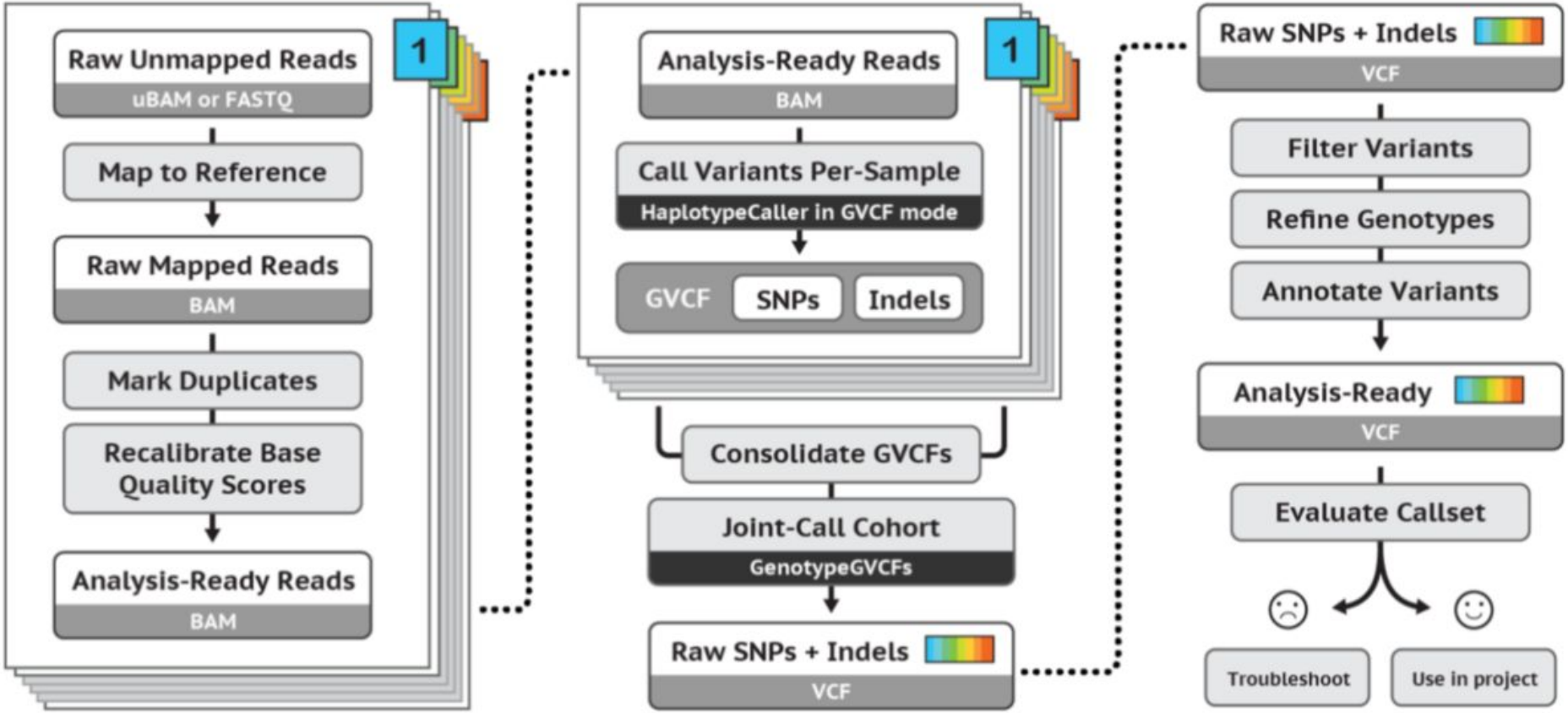
Read group(s)

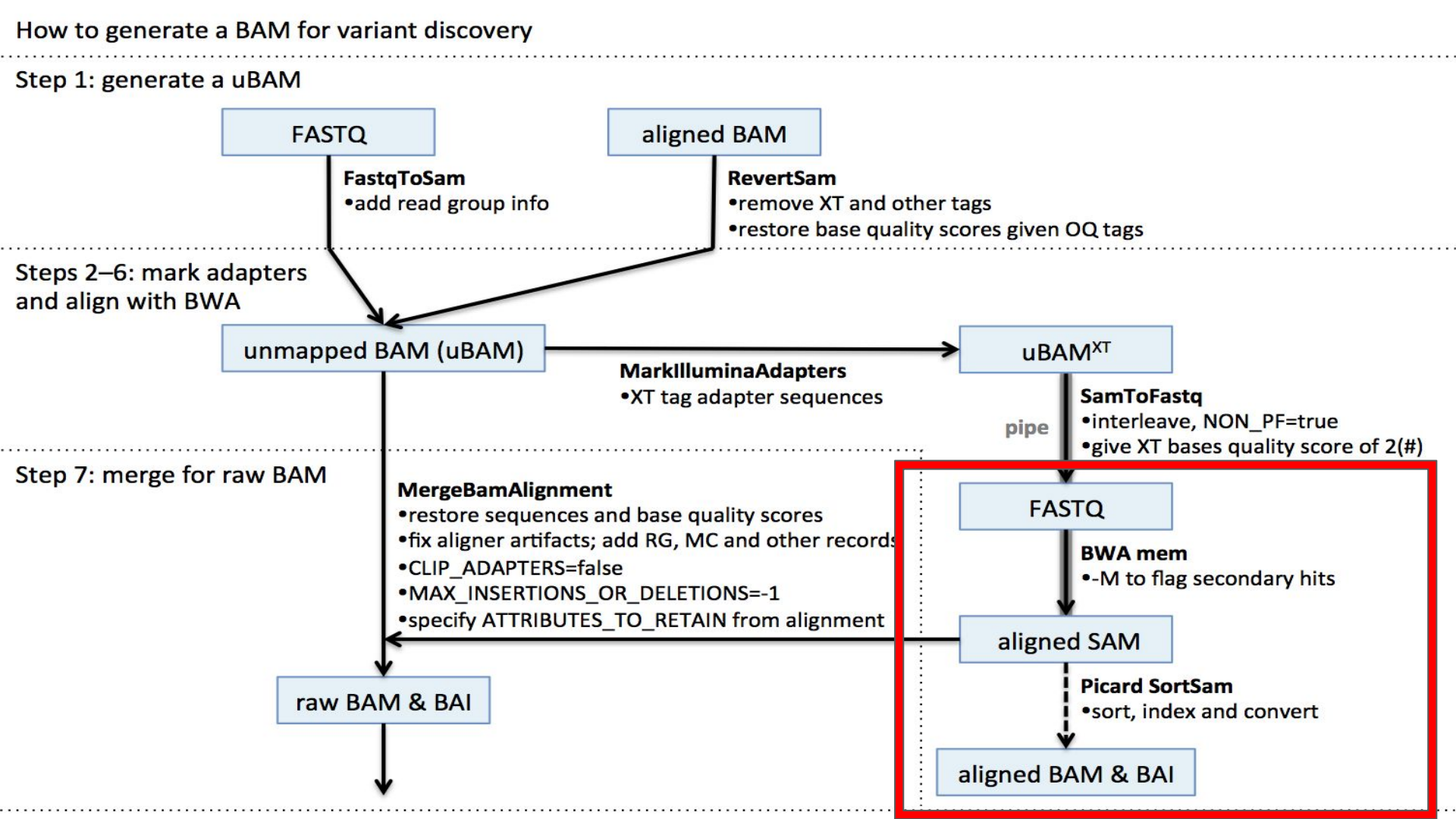
RECORDS containing structured read information (1 line per read/record)



- Added mapping info summarizes **position**, **quality**, and **structure** for each **read**
- Mate information points to the read from the other end of the molecule (other in a pair)

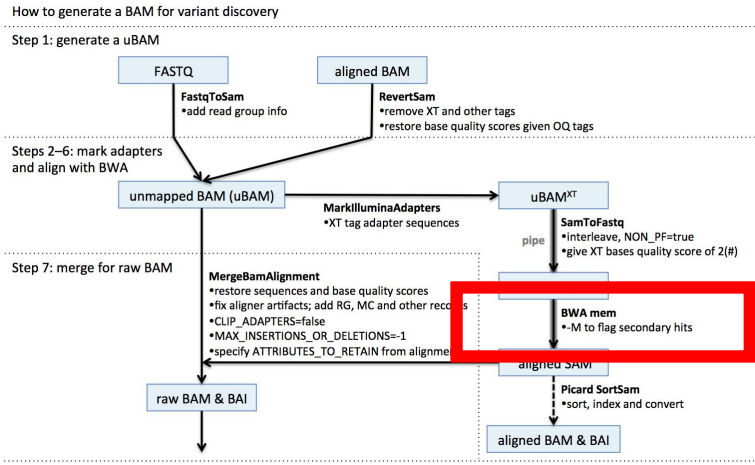
```
SRR062641.1136801       0       1       41692   0       100M    *       0       A
TTGGTTCAAAGTATAACATGTTAAAGCACAGAGCCCCAACTCTGAAAAGTACCATCCCTAAATTGGCATTTAGTTGCACCT
TTATATCCACCTTTAAAA      3<=;<;?;<>><<>>;>;;==<?>><<<=;><><<;=<<=:<<==<===<;;;<<<;
;;<;;<:=;<;<<==;<<9<<;;<<:9;<;;*<;9:5:<:-9<     MD:Z:100        PG:Z:MarkDuplicat
es      RG:Z:4  NM:i:0  UQ:i:0  AS:i:100        XS:i:100
SRR062641.16863150      0       1       66399   9       20S51M3I15M1D11M        T
ATTATATAATTTTATTTATTTATATAATAATATATATTTTATTATATAAATATATATTATATTATATAATATAATATATAT
TAATATAAATATATATTA      3<<=<<<<<=<===<<==<<==<<<<<=<<=<<<<<<<===<<=<<<<<==<=<<<<
<=<<<<=<<<<<<<<<<=<<<<<<<=<=<<<<<<<<<<<<=<     MD:Z:10T0A45T3T4^T11    PG:Z:Mark
Duplicates      RG:Z:4  NM:i:8  UQ:i:148        AS:i:44 XS:i:38
```


Raw Unmapped Reads
uBAM or FASTQ
1
Map to Reference
Raw Mapped Reads
BAM
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM
Analysis-Ready Reads
BAM
1
Call Variants Per-Sample
HaplotypeCaller in GVCF mode
GVCF
SNPs
Indels
Consolidate GVCFs
Joint-Call Cohort
GenotypeGVCFs
Raw SNPs + Indels
VCF
Raw SNPs + Indels
VCF
Filter Variants
Refine Genotypes
Annotate Variants
Analysis-Ready
VCF
Evaluate Callset
Troubleshoot
Use in project


How to generate a BAM for variant discovery
Step 1: generate a uBAM
FASTQ
aligned BAM
FastqToSam
•add read group info
RevertSam
•remove XT and other tags
•restore base quality scores given OQ tags
Steps 2–6: mark adapters and align with BWA
unmapped BAM (uBAM)
uBAM^XT
MarkIlluminaAdapters
•XT tag adapter sequences
pipe
SamToFastq
•interleave, NON_PF=true
•give XT bases quality score of 2(#)
Step 7: merge for raw BAM
MergeBamAlignment
•restore sequences and base quality scores
•fix aligner artifacts; add RG, MC and other records
•CLIP_ADAPTERS=false
•MAX_INSERTIONS_OR_DELETIONS=-1
•specify ATTRIBUTES_TO_RETAIN from alignment
FASTQ
BWA mem
•-M to flag secondary hits
aligned SAM
Picard SortSam
•sort, index and convert
raw BAM & BAI
aligned BAM & BAI

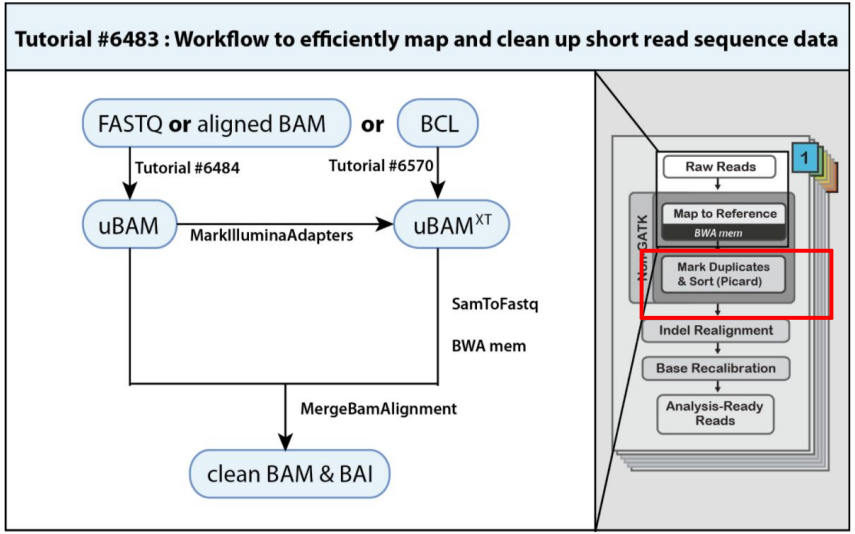
2. FASTQ to BAM - Alignment

- Use BWA mem and reference genome
- bwa mem -M -t 7 -p human_g1k_v37.fasta HG_input.fq > HG_aligned.sam
- **D and E** in the Code document



2. FASTQ to BAM - MarkDuplicates and SortSam

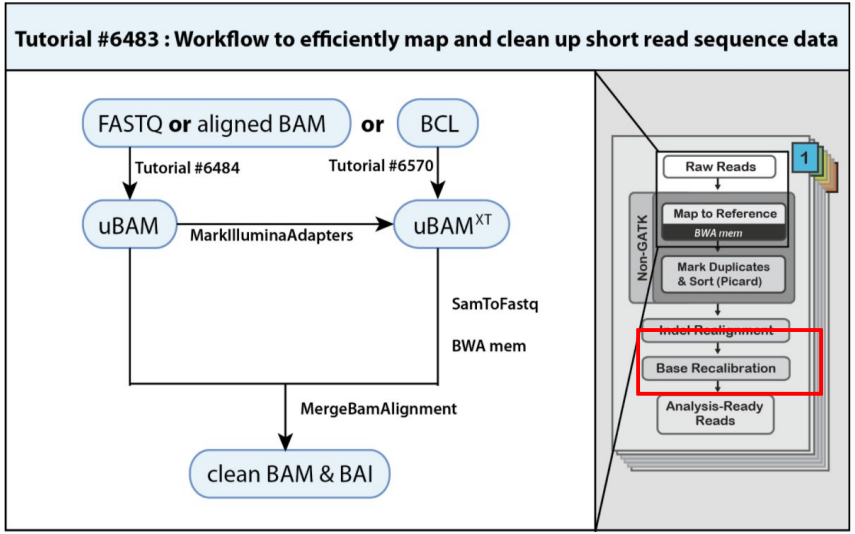
- Finding duplicates is important for calling accuracy and expression level accuracy (RNA)
- java -jar picard.jar MarkDuplicates I=HG_preprocessed.bam O=HG_dedup.bam M=HG_dedup.metrics.txt
- java -jar picard.jar SortSam I=HG_dedup.bam O=HG_sorted.bam SO=coordinate



2. FASTQ to BAM - Base Recalibration

- Indel realignment is deprecated now (haplotypecaller does that)
- gatk --java-options '-Xmx16g' BaseRecalibrator -I HG_sorted.bam -R human_g1k_v37.fasta --known-sites ALL.wgs.mergedSV.v8.20130502.svs.genotypes.vcf.gz -O HG.recal_data.table
- gatk --java-options '-Xmx16g' ApplyBQSR -I HG_sorted.bam -R human_g1k_v37.fasta --bqsr-recal-file HG.recal_data.table -O HG.fin.bam

DONE!

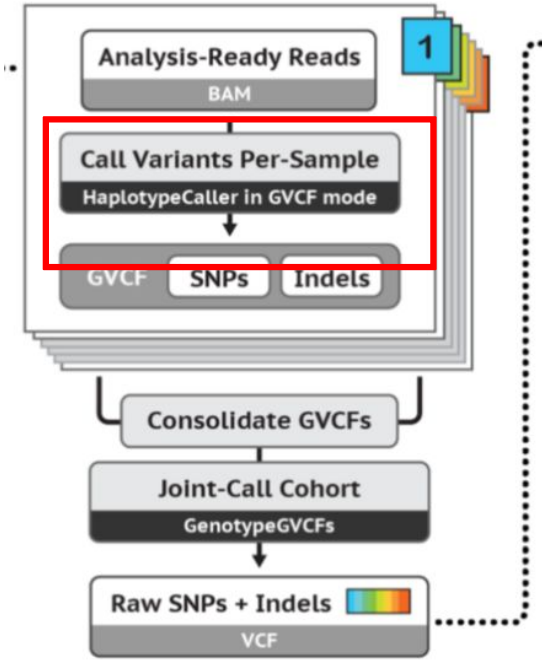


3. BAM to VCF

a. gvcf using haplotype caller
b. combining gvcf into vcf

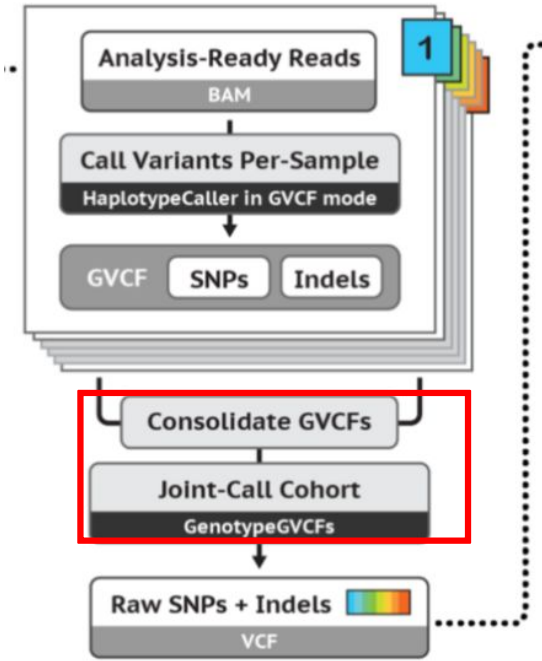
3. BAM to VCF - making gvcf file with reference and bam

gatk --java-options "-Xms4g" **HaplotypeCaller** -R /media/leelabsg_storage01/1000G/human_g1k_v37.fasta -I /home/lee7801/DATA/1000G/HG00096.chrom20.ILLUMINA.bwa.GBR.exome.20120522.bam -L 20 -ERC GVCF -O sample01_20.g.vcf



3. BAM to VCF - combine gvcf and convert to vcf

- gatk CombineGVCFs -R /media/leelabsg_storage01/1000G/human_g1k_v37.fasta --variant sample01_20.g.vcf --variant sample02_20.g.vcf --variant sample03_20.g.vcf -O sample_all.g.vcf.gz
- gatk --java-options "-Xmx4g" GenotypeGVCFs -R /media/leelabsg_storage01/1000G/human_g1k_v37.fasta -V sample_all.g.vcf.gz -O sample_all_fin.vcf



Reference

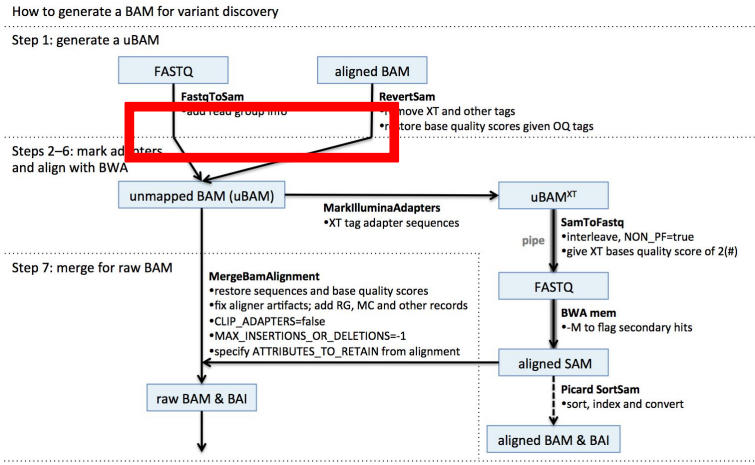
- https://drive.google.com/drive/folders/1Nh73FzKde203gUoxyR9CmTd1EcVDMCI5
- https://www.internationalgenome.org/faq/which-reference-assembly-do-you-use/
- https://www.biostars.org/p/174343/
- https://github.com/broadgsa/gatk-protected/blob/master/doc_archive/deprecated/%5BHow_to%5D_Generate_a_BAM_for_variant_discovery_(long).md#step7
- https://gatk.broadinstitute.org/hc/en-us/articles/360035535912-Data-pre-processing-for-variant-discovery
- https://gatk.broadinstitute.org/hc/en-us/articles/360035890531
- https://2wordspm.com/2019/09/23/ngs-dna-seq-pipeline-gatk-best-practice-code-part1-fastq-to-bam/

Appendix. FASTQ to BAM - unmapped bam file

- GATK regards fastq file inferior to uBAM file format because it lacks ability to store metadata.
- https://gatk.broadinstitute.org/hc/en-us/articles/360039568932--How-to-Map-and-clean-up-short-read-sequence-data-efficiently#step3D
- FASTQ : picard FastqToSam
- Aligned bam file : picard RevertSam

For the fastq, add read group

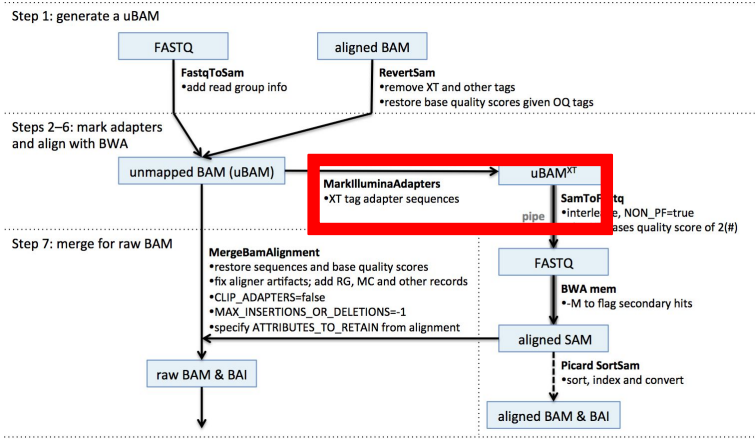
https://gatk.broadinstitute.org/hc/en-us/articles/360037226472-AddOrReplaceReadGroups-Picard-



Appendix. FASTQ to BAM - adapter sequence and conversion

- Use picard MarkIlluminaAdapters and SamToFastq
- Convert the bam file to fastq to **remove adapter portion of the sequence** and we use this fastq file as input

XT:i:88
3KJMPPPPPPPOROPQQRROQOQROOQOPQROQOPPPQROOPPOQSPPPPRSPRQPPPPPPQQMMPPOPPPOOPLRSHROMJOPORH############
3KJMPPPPPPPOROPQQRROQOQROOQOPQROQOPPPQROOPPOQSPPPPRSPRQPPPPPPQQMMPPOPPPOOPLRSHROMJOPORHOINOQPOHNHFFJ



Appendix. FASTQ to BAM - MergeBamAlignment

- Use picard MergeBamAlignment
- Merge metadata from the unmapped bam file and alignment information from the aligned sam file
- **F** in Code document