Reads Alignment and Variant Calling

CB2-201 – Computational Biology and Bioinformatics
February 22, 2016

Emidio Capriotti
http://biofold.org/

Biomolecules Folding and Disease

Institute for Mathematical Modeling
of Biological Systems
Department of Biology

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Genome Analysis

Alignment is different from Assembly:

- Alignment aims to find the best matches of a particular read to a reference genome

- Assembly finds the best overlaps among the reads to determine the most likely genome

Mapping

Given the level of variability across individuals it is expected that an high fraction of reads will not map.

The complete human pan-genome will require additional 19-40 Mb of novel sequences. *Nature Biotechnology 28, 57–63 (2010)*

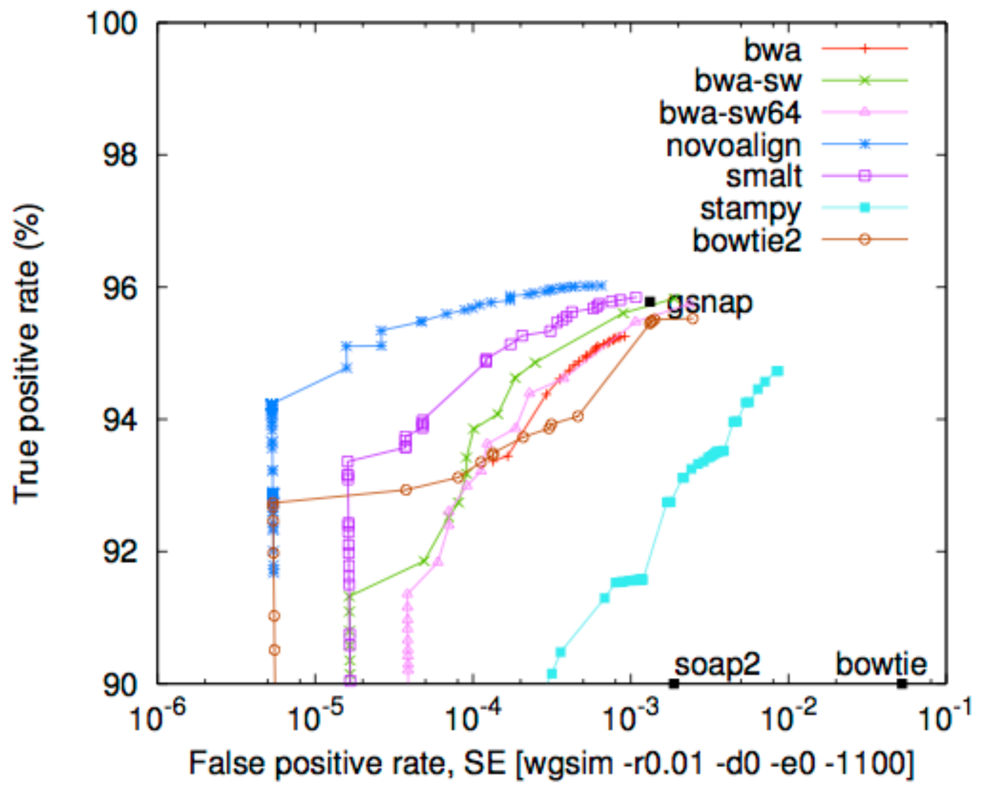
The Alignment

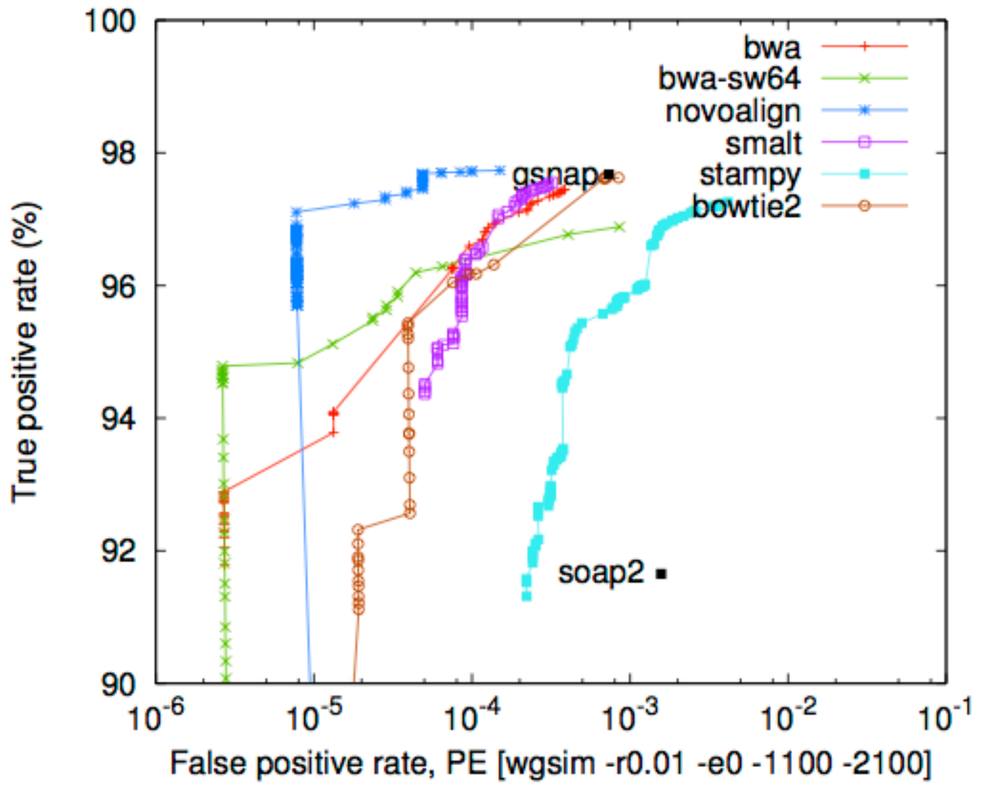
The main limitation is the computational time.
the best method should have a good balance between cpu and memory usage, speed and accuracy.

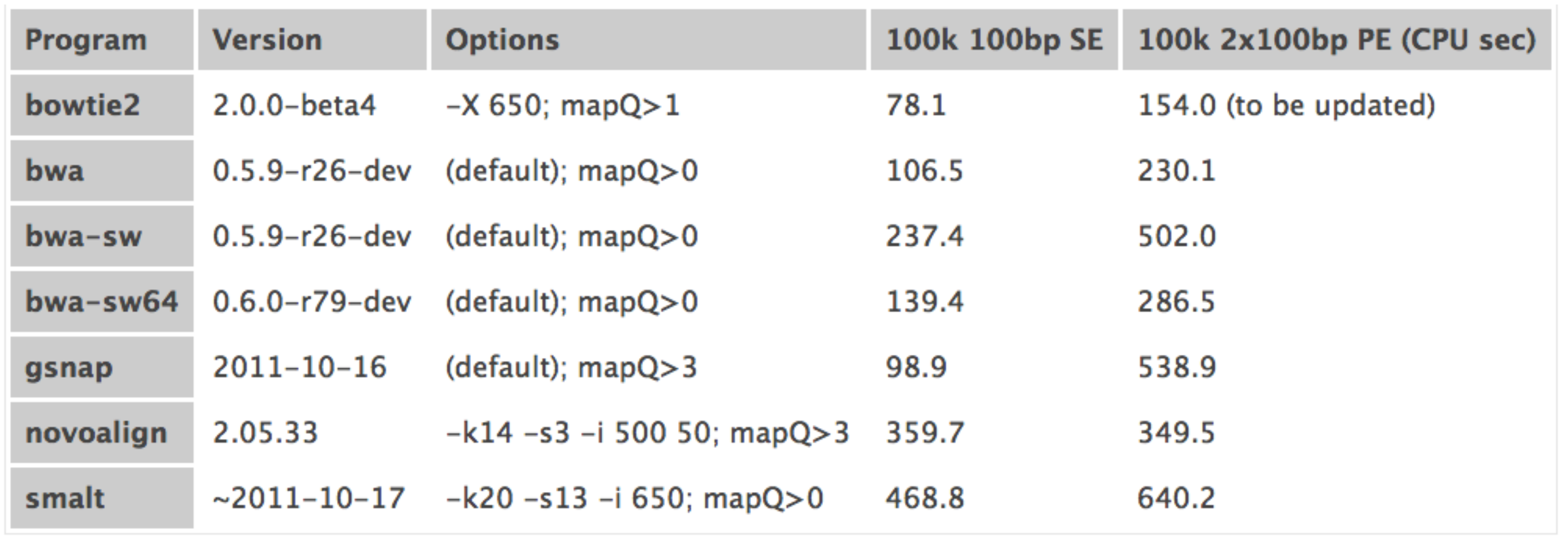
The most popular methods (BWA and Bowtie) are based on suffix array: a sorted array of all suffixes of a string.

BWA and other aligners such as Bowtie use an implementation of the Burrows–Wheeler transform algorithm which is a technique for data compression.

Testing methods







Program	Version	Options	100k 100bp SE	100k 2x100bp PE (CPU sec)
bowtie2	2.0.0-beta4	-X 650; mapQ>1	78.1	154.0 (to be updated)
bwa	0.5.9-r26-dev	(default); mapQ>0	106.5	230.1
bwa-sw	0.5.9-r26-dev	(default); mapQ>0	237.4	502.0
bwa-sw64	0.6.0-r79-dev	(default); mapQ>0	139.4	286.5
gsnap	2011-10-16	(default); mapQ>3	98.9	538.9
novoalign	2.05.33	-k14 -s3 -i 500 50; mapQ>3	359.7	349.5
smalt	~2011-10-17	-k20 -s13 -i 650; mapQ>0	468.8	640.2

http://lh3lh3.users.sourceforge.net/alnROC.shtml

Genome Indexing

The first step consist in the creation of an indexed reference genome.

```
> bwa index $db.fasta
```

two useful option:

```
-p database prefix
-a indexing algorithm
     “bwtsw” for large genome (> 50,000,000 BP)
     “is” for smaller genomes
```

Reference from different institutions
ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/2.8/hg19/

Other Indexing

GATK requires two specific index files with extension fai and dict.

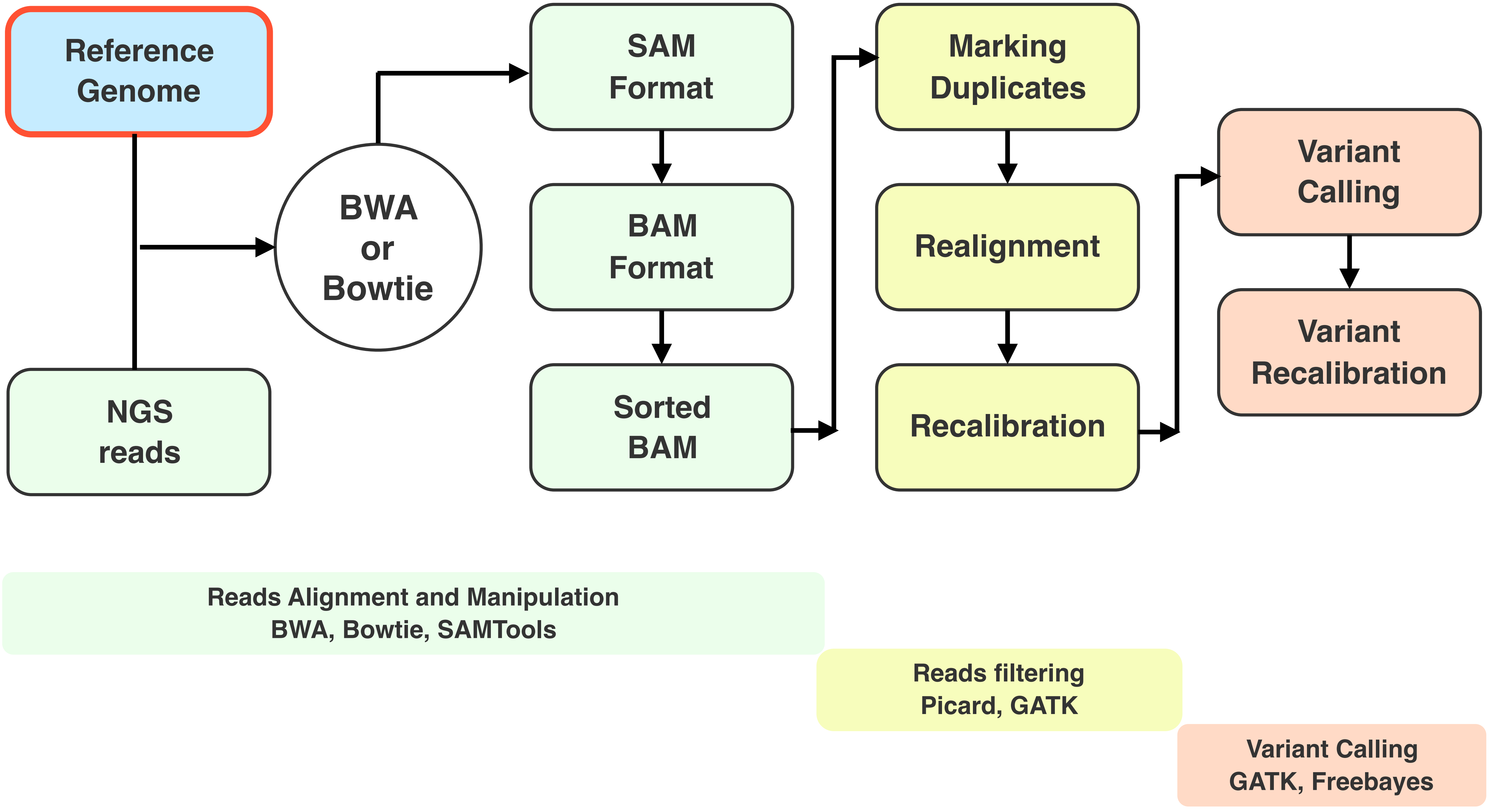
The fai file is generated by *samtools*

```
> samtools faidx $db.fasta
```

The dict file is generated by *picard*

```
>java -jar CreateSequenceDictionary.jar \
    REFERENCE=$db.fasta OUTPUT=$db.dict
```

Variant Calling Steps



Alignments of the reads

The alignment is generated using bwa aln

```
> bwa aln -t [opts] $db $file.fastq
```

two useful option:

```
-f output file
-t number of threads
```

If you are have analyzing pair-end sequencing you need to repeat the alignment for both *fastq* files.

Generate the SAM file

The SAM file is generated using bwa samse (single-end) or sampe (pair-end)

```
> bwa samse $db $file.sai $file.fastq
> bwa sampe $db $file1.sai $file2.sai $file1.fastq $file2.fastq
```

two useful option:

```
-f output file
-r group_info
     for example
     @RG\tID:. .\tLB:. .\tSM:. .\tPL:ILLUMINA
```

SAM to BAM

To save space and make all the process faster we can convert the SAM file to BAM using *samtools*

```
> samtools view -bS $file.sam > $file.bam
```

useful option:

```
-b output BAM

-S input SAM

-T reference genome if header is missing
```

Samtools functions

Samtools can be used to perform several tasks:

Sorting BAM file

```
> samtools sort $file.bam -o $file.sorted
```

Create an index

```
> samtools index $file.bam $file.bai
```

Filtering out unmapped reads

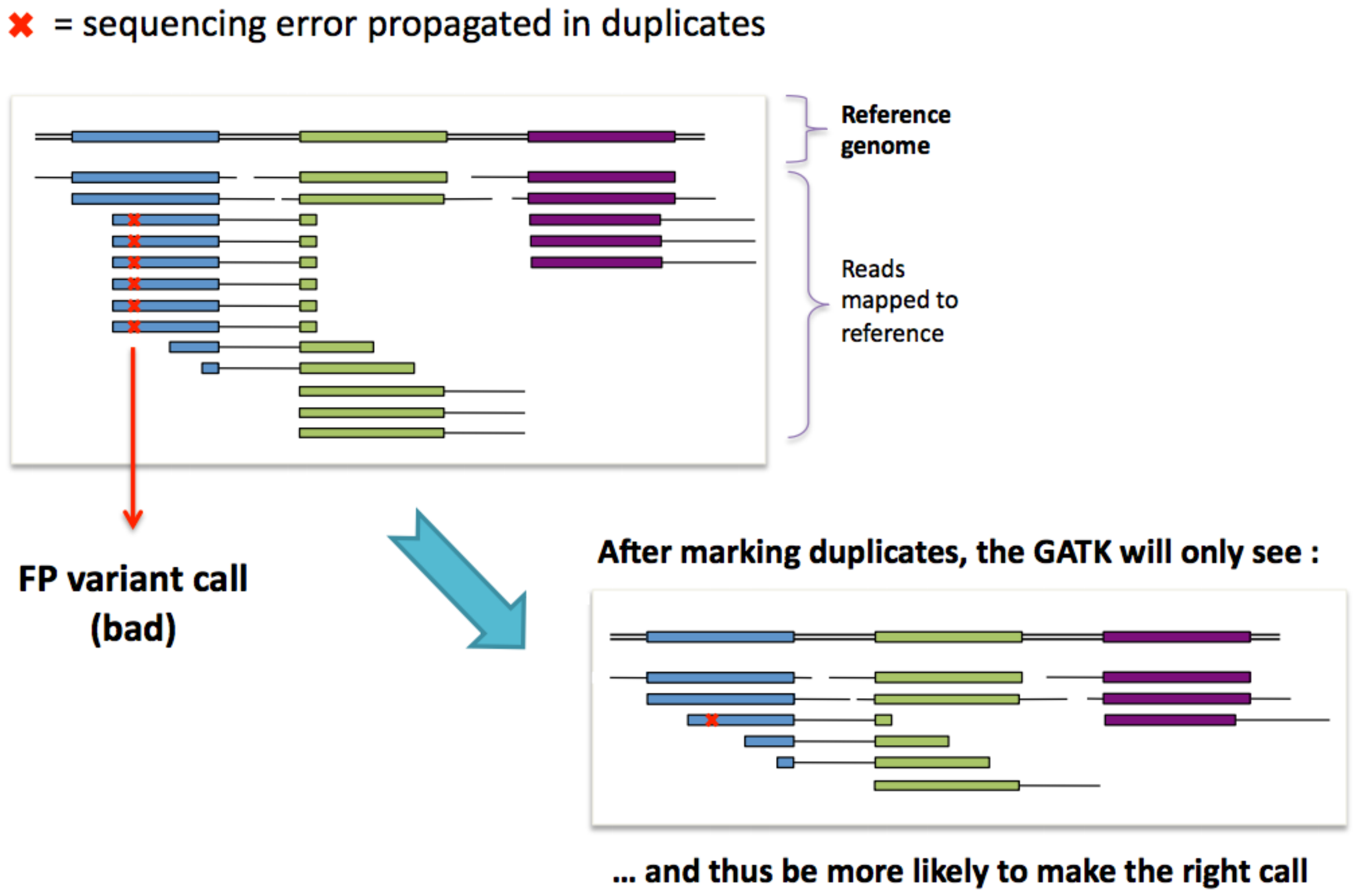
```
> samtools view -h -F 4 $file.bam $file.mapped.bam
```

Select properly paired reads

```
> samtools view -h -f 0X0002 $file.bam $file.paired.bam
```

Optical Duplication

Optical duplicates are due to a read being read twice. The number of the duplicates depends on the depth of the sequencing, the library and sequencing technology.

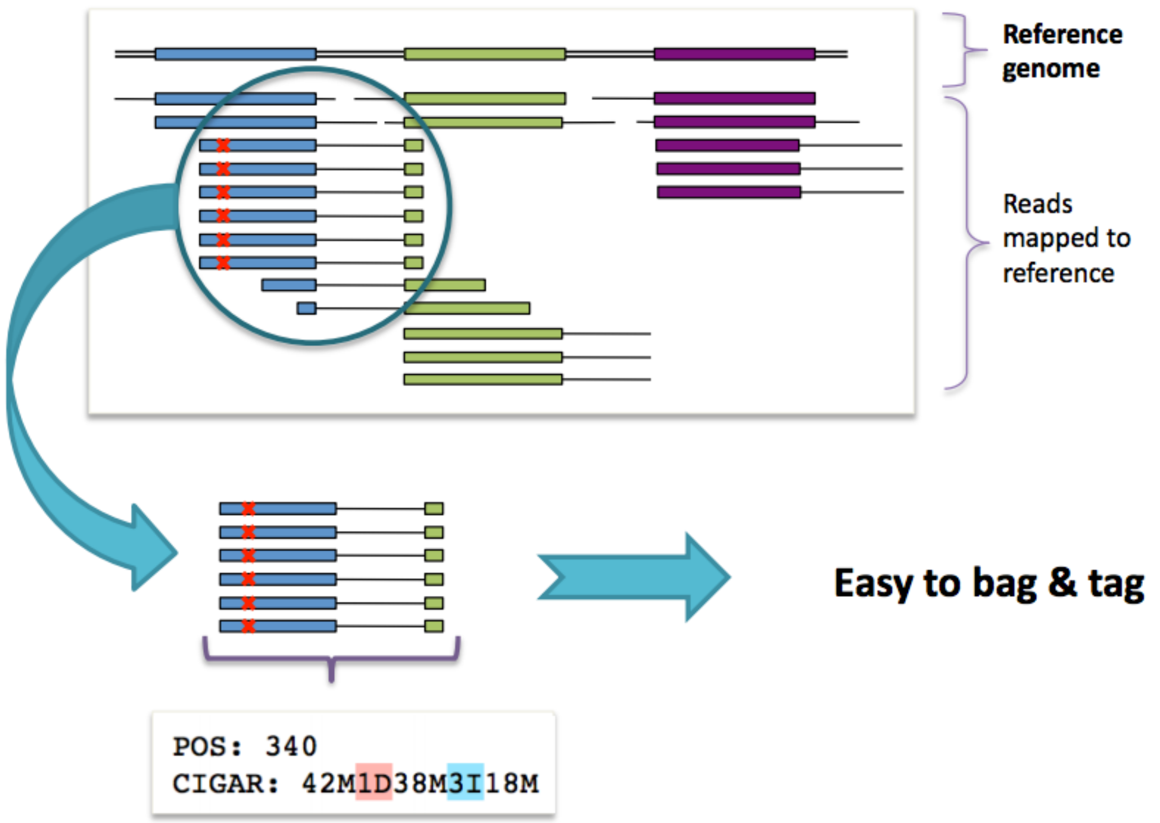

✖ = sequencing error propagated in duplicates
Reference genome
Reads mapped to reference
FP variant call (bad)
After marking duplicates, the GATK will only see :
... and thus be more likely to make the right call

https://www.broadinstitute.org/gatk/

Picard

Duplication can be detect comparing the
CIGAR (Short way to represent alignment with reference sequence).
Picard is used to mark the duplicated reads.

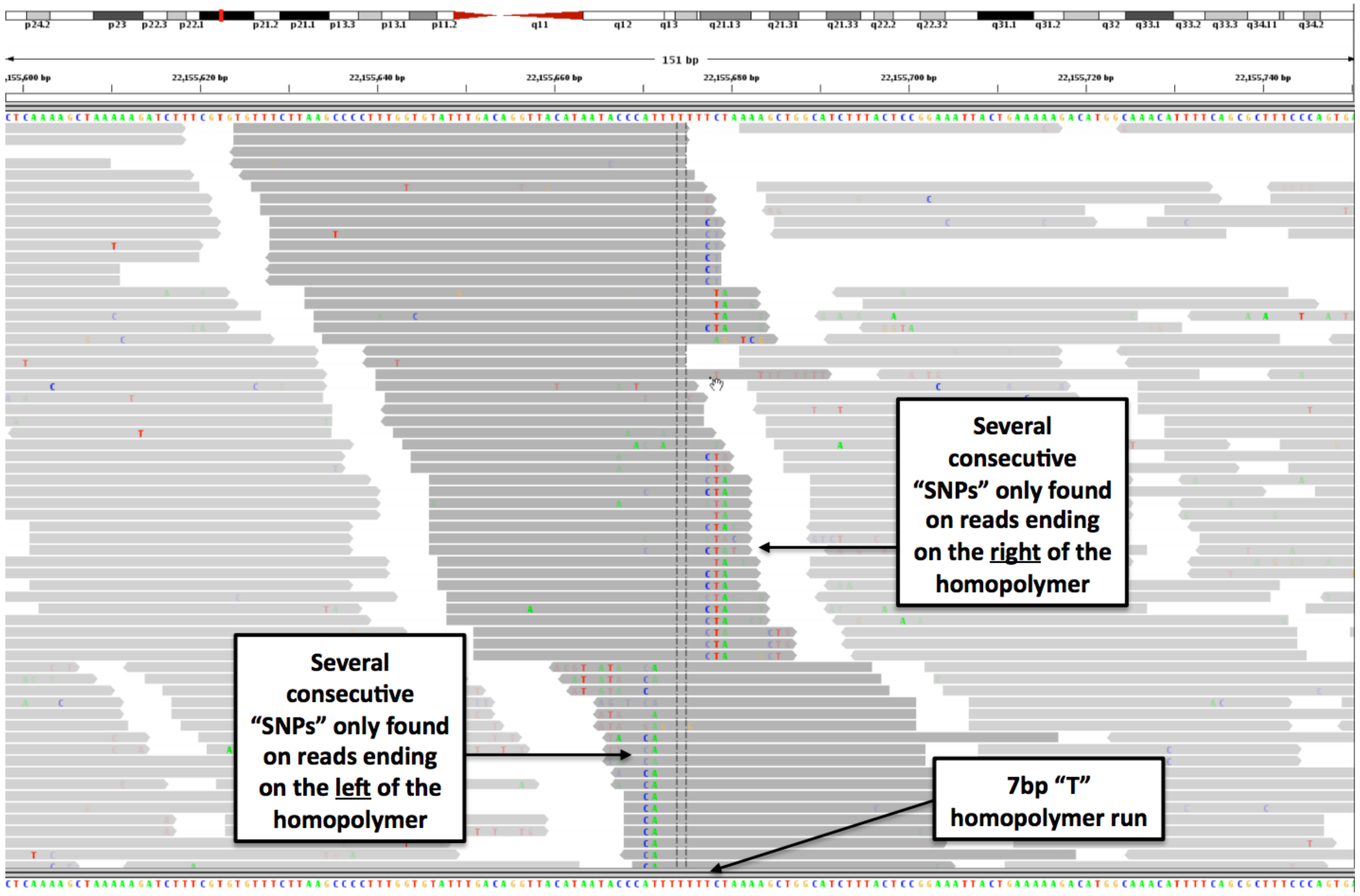
```
# Mark duplicate

java -Xmx4g -Djava.io.tmpdir=/tmp -jar MarkDuplicates.jar \
INPUT=$file.bam OUTPUT=$file.marked.bam METRICS_FILE=metrics \
CREATE_INDEX=true  VALIDATION_STRINGENCY=LENIENT
```



https://www.broadinstitute.org/gatk/

Indel realignment

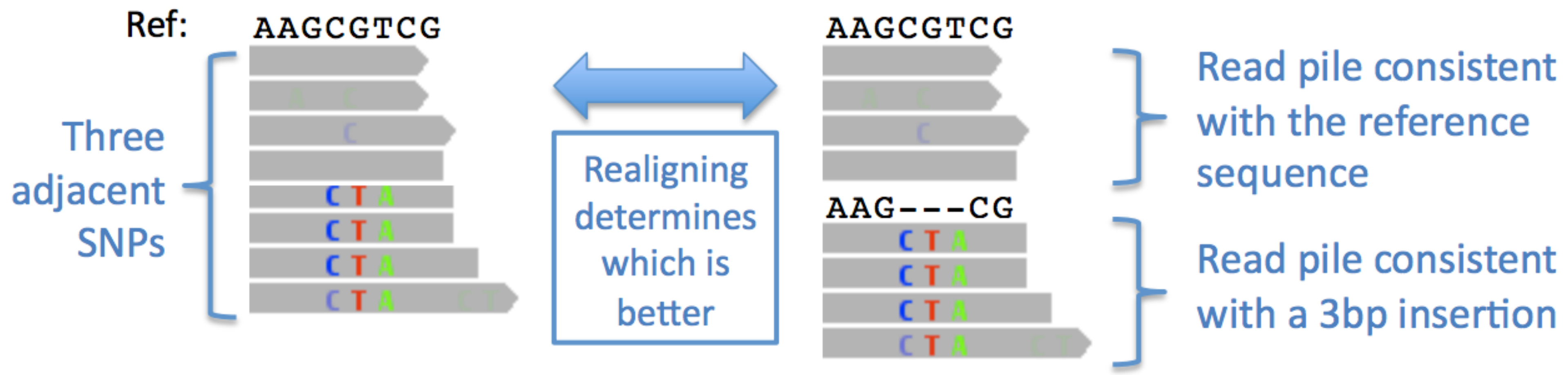
The mapping of indels especially in regions near to the ends can be seen as mismatches. Consecutive variants close to the ends are suspicious.


151 bp
Several consecutive "SNPs" only found on reads ending on the right of the homopolymer
Several consecutive "SNPs" only found on reads ending on the left of the homopolymer
7bp "T" homopolymer run

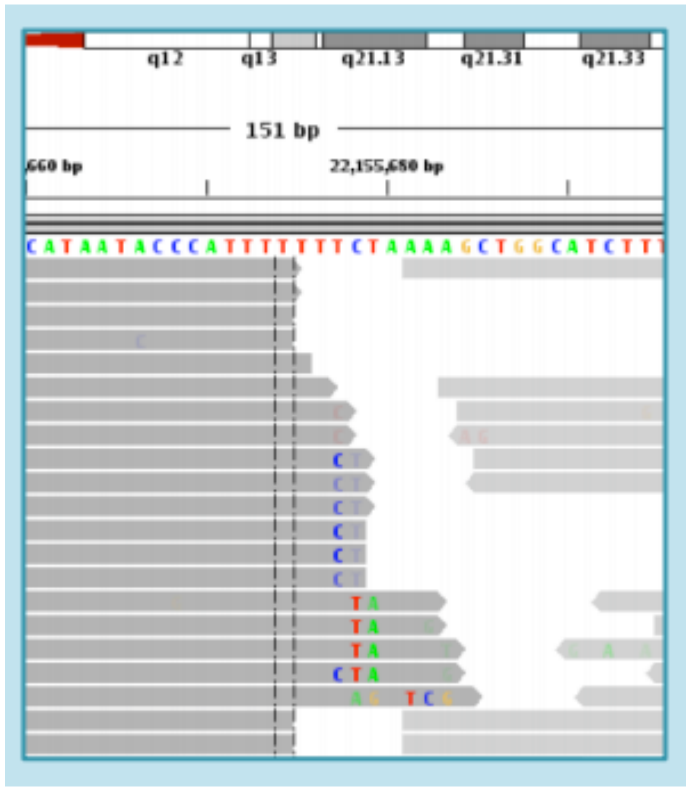
https://www.broadinstitute.org/gatk/

Better Alignment

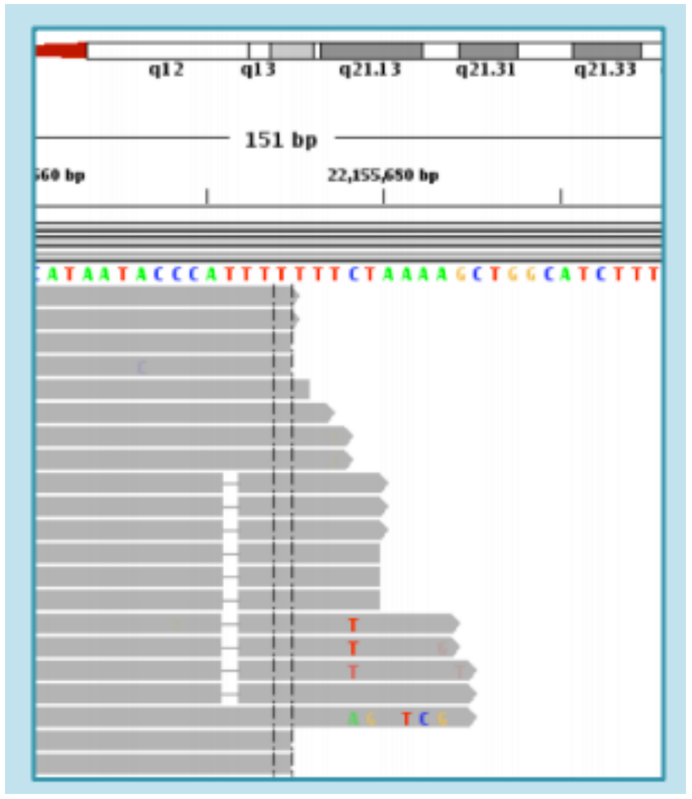
The realignment around the indels improve the quality of the alignment


Ref:
AAGCGTCG
Three adjacent SNPs
Realigning determines which is better
AAGCGTCG
AAG---CG
Read pile consistent with the reference sequence
Read pile consistent with a 3bp insertion

RealignerTargetCreator



IndelRealigner



https://www.broadinstitute.org/gatk/

Realignment

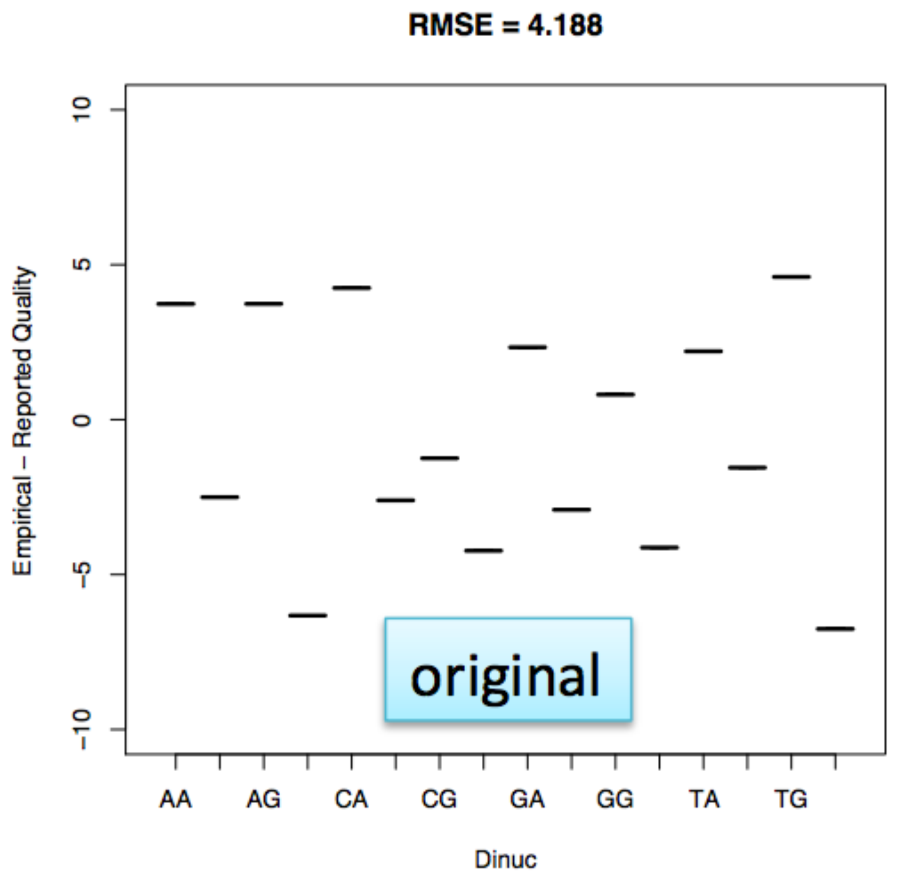
After marking the duplicated reads GATK the alignment is recalculated to improve the mach to the indels.

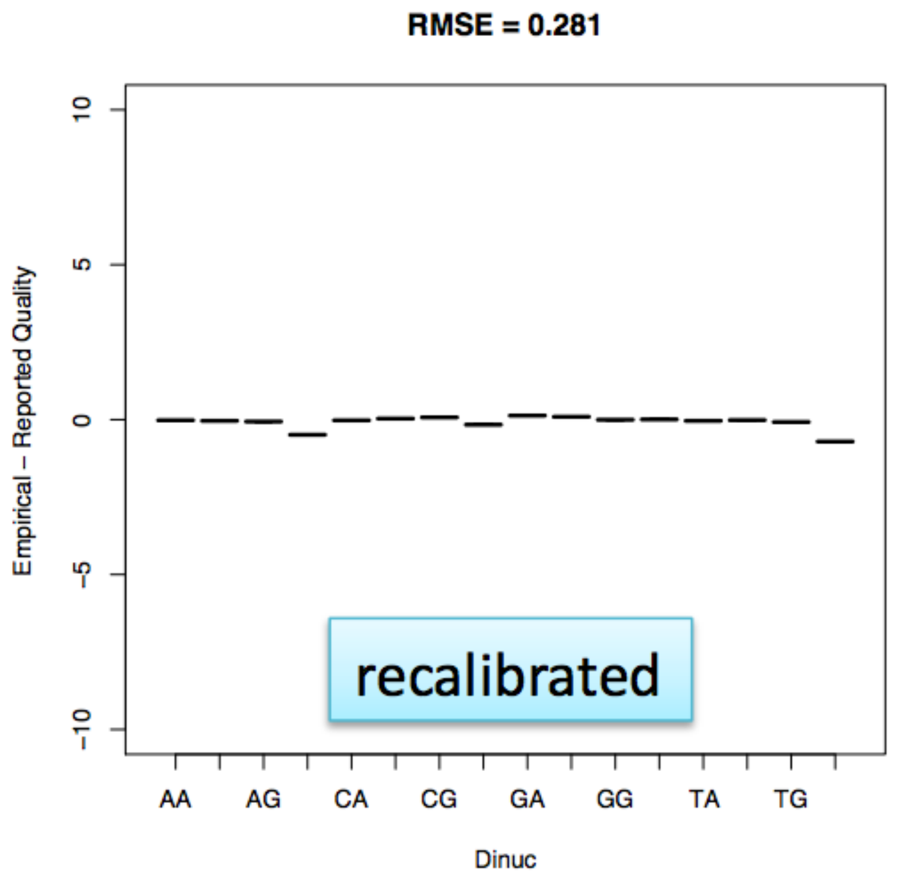
```
# Local realignment

java -Xmx4g -jar GenomeAnalysisTK.jar -T RealignerTargetCreator \
-R $db.fa -o $file.bam.list  -I $file.marked.bam

java -Xmx4g -Djava.io.tmpdir=/tmp -jar GenomeAnalysisTK.jar \
-I $file.marked.bam -R $genome -T IndelRealigner \
-targetIntervals $file.bam.list -o $file.marked.realigned.bam
```

```
# Picard realignment

java -Djava.io.tmpdir=/tmp/flx-auswerter -jar FixMateInformation.jar \
INPUT=$file.marked.realigned.bam OUTPUT=$file.marked.realigned.fixed.bam \
SO=coordinate VALIDATION_STRINGENCY=LENIENT CREATE_INDEX=true
```

Quality Score

The quality score are critical for the downstream analysis. This score depends on the nucleotide context.





https://www.broadinstitute.org/gatk/

The recalibration step

After removing the duplicated reads scores need to be recalibrated using GATK.
The recalibration is detected calculating the covariation among nucleotide features.

```
# Recalibration
java -Xmx4g -jar GenomeAnalysisTK.jar -T BaseRecalibrator  \
-R $db.fa  -I $file.marked.realigned.fixed.bam -knownSites $dbsnp -o \
$output.recal_data.csv

java -jar GenomeAnalysisTK.jar -T PrintReads \
-R $db.fa  -I $file.marked.realigned.fixed.bam \
-BQSR $file.recal_data.csv -o $file.marked.realigned.fixed.recal.bam
```

The Variant Calling

There are two possible options: UnifiedGenotyper and HyplotypeCaller

```
# Variant calling

java -Xmx4g -jar GenomeAnalysisTK.jar -glm BOTH -R $db.fa \
-T UnifiedGenotyper -I $file.marked.realigned.fixed.recal.bam \
-D $dbsnp -o $file.vcf -metrics snps.metrics  \
-stand_call_conf 50.0 -stand_emit_conf 10.0 -dcov 1000 -A AlleleBalance
```

or

```
java -Xmx4g -jar GenomeAnalysisTK.jar -T HaplotypeCaller -R $db.fa \
-I $file.marked.realigned.fixed.recal.bam --emitRefConfidence GVCF \
—variant_index_type LINEAR --variant_index_parameter 128000 --dbsnp $dbsnp \
-o $output.recalibrated.vcf
```

UnifiedGenotyper is faster and HyplotypeCaller is more accurate on the detection of indels.

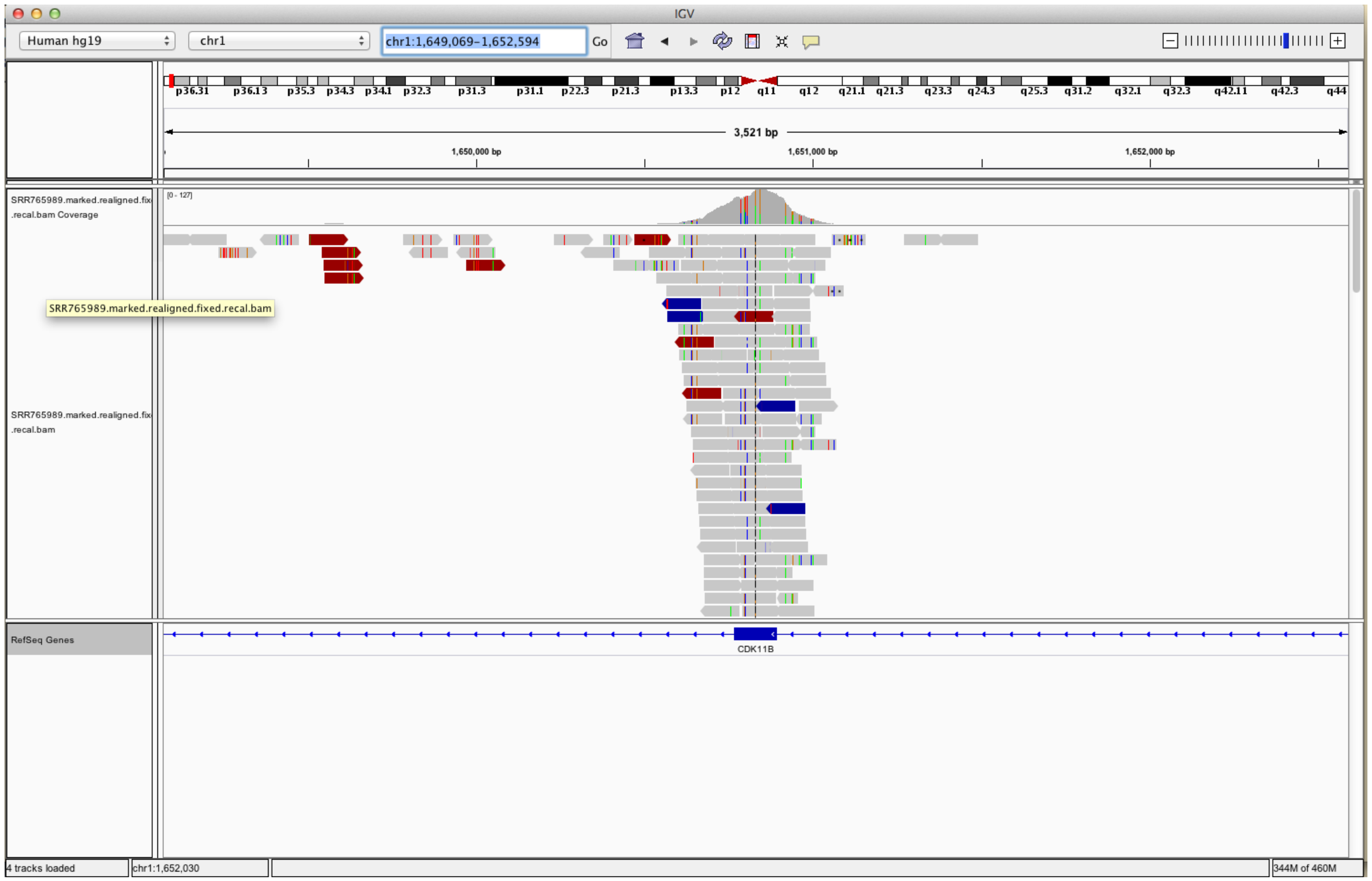
Improve Variant Calling

The final step consist in recalibration and filtering. The letter are based on previously known mutation events.

```
# Variant Quality Score Recalibration
java -jar GenomeAnalysisTK.jar –T VariantRecalibrator\
–R     $db.fa –input $file.vcf     \
–resource: {dbsnp, 1000Genomes, Haplotype }    \
–an QD –an MQ –an HaplotypeScore {…}    \
–mode  SNP –recalFile $file.snps.recal   \
–tranchesFile $file.recalibrated.tranches


java -jar GenomeAnalysisTK.jar -T ApplyRecalibration \
-R $db.fa -input $file.vcf  -mode SNP\
-recalFile $file.snps.recal -tranchesFile raw.SNPs.tranches \
-o $file.recalibrated.vcf -ts_filter_level    99.0


# Variant Filtering

java -Xmx4g -jar GenomeAnalysisTK.jar -R $db.fa \
-T VariantFiltration -V $file.recalibrated.vcf \
-o $file.recalibrated.filtered.vcf --clusterWindowSize 10 \
--filterExpression “some filter --filterName “filter_name”
```

Visualization

Broad Institute have developed the Integrative Genome Viewer (IGV) for the visualization of genomic data from different sources of data.



http://cmb.path.uab.edu/training/docs/CB2-201-2015/IGV_2.3.40.zip

For more details

Samtools

http://www.htslib.org/doc/

GATK Guide

https://www.broadinstitute.org/gatk/guide/

Best practices for variant calling with GATK

http://www.broadinstitute.org/partnerships/education/broade/best-practices-variant-calling-gatk

IGV

http://www.broadinstitute.org/igv/