Séquençage d'exome pour la recherche de variants génomiques en oncologie clinique

Avec des diapos, données & scripts R de:
Yannick Boursin, IGR
Bastien Job, IGR

UNIVERSITÉ PARIS SUD

GUSTAVE ROUSSY
CANCER CAMPUS GRAND PARIS

Génétique constitutionnelle

At hospital

Blood sample → Sequence gene panel → Look for specific alteration (BRCA)

Research

Genotype or Sequence → Compare disease and healthy cohorts

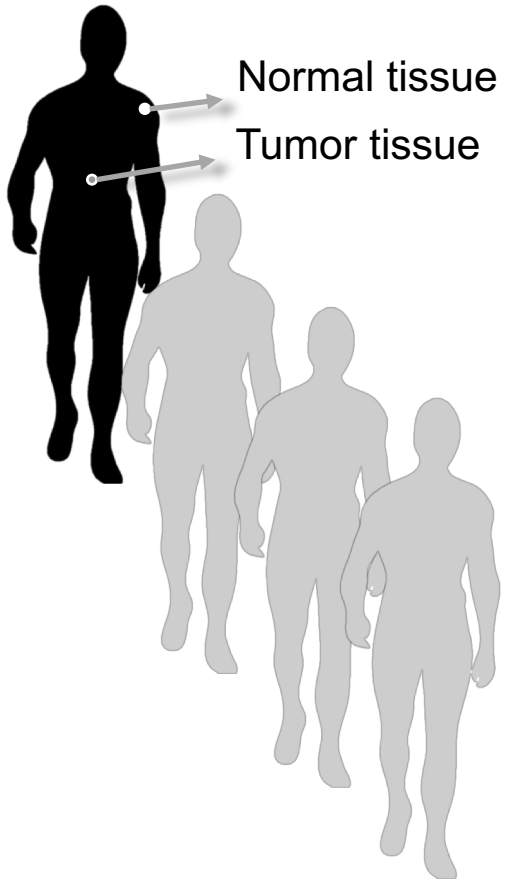
GWAS studies, 1000 Genome Project…

NGS dans le diagnostic de génétique familiale

- BRCA1/2 (breast/ovary cancer)
- XPC, XPV.. (melanoma)
- ERCC1 (colorectal cancer)

Génétique somatique

Finding somatic mutations in the tumor genome

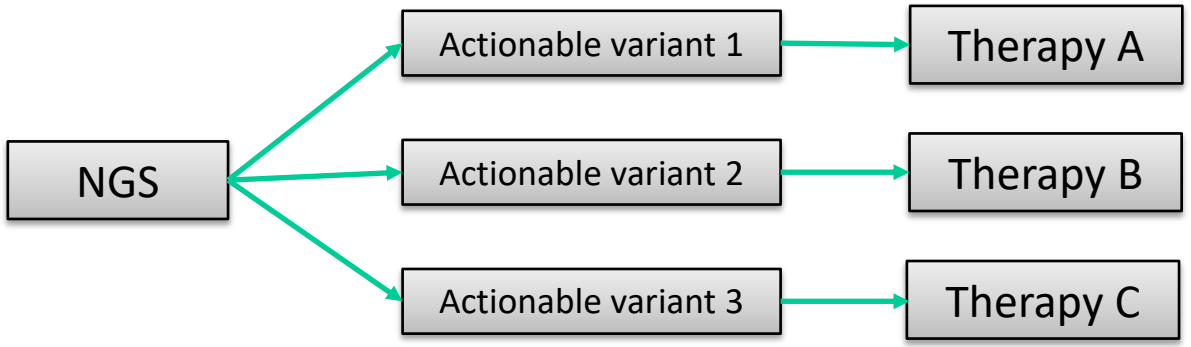


Normal tissue (blood) → Extract and sequence DNA

Tumor tissue → Look for mutations, structural variations

Repeat for several patients → Look for recurrent events

NGS for precision medicine



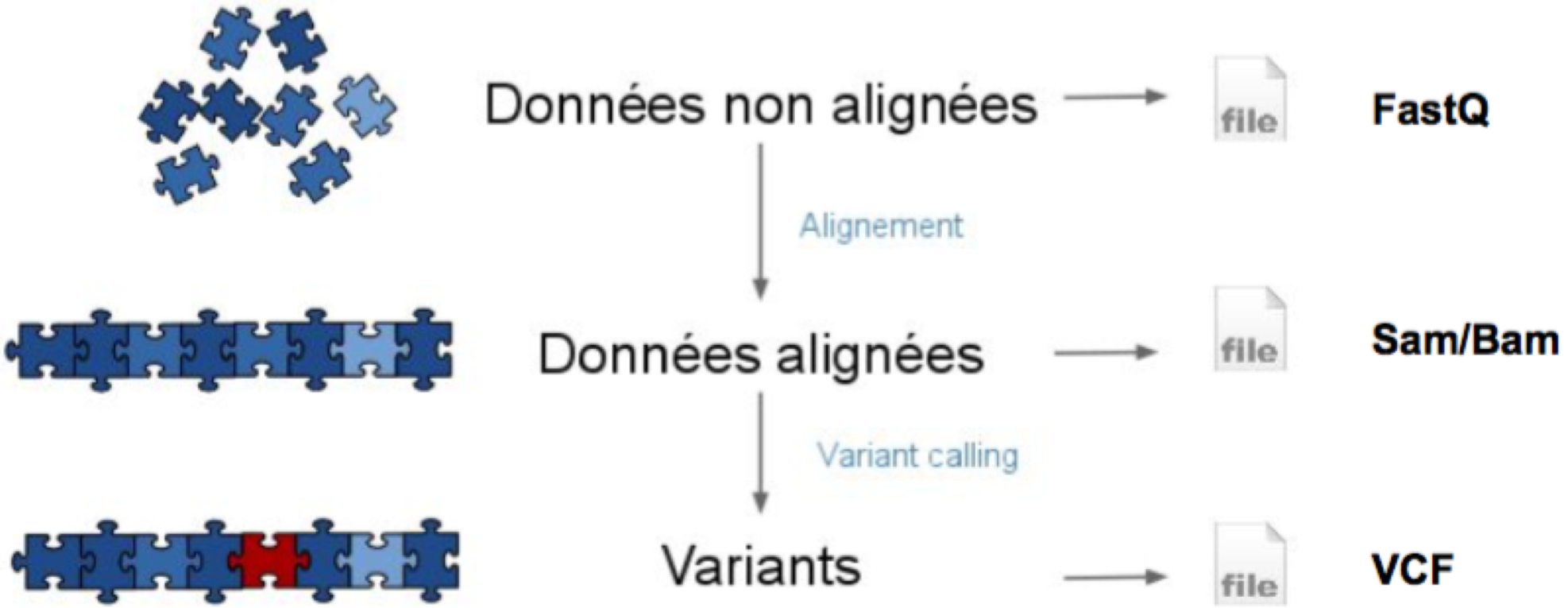
– Clinical trials: MOSCATO (GR), SAFIR (GR), SHIVA (Curie), …

Drugs: Ipilimumab (anti-CTLA4), Nivolumab (anti-PD1), Trastuzumab (anti-HER2), Cetuximab (anti-EGFR)

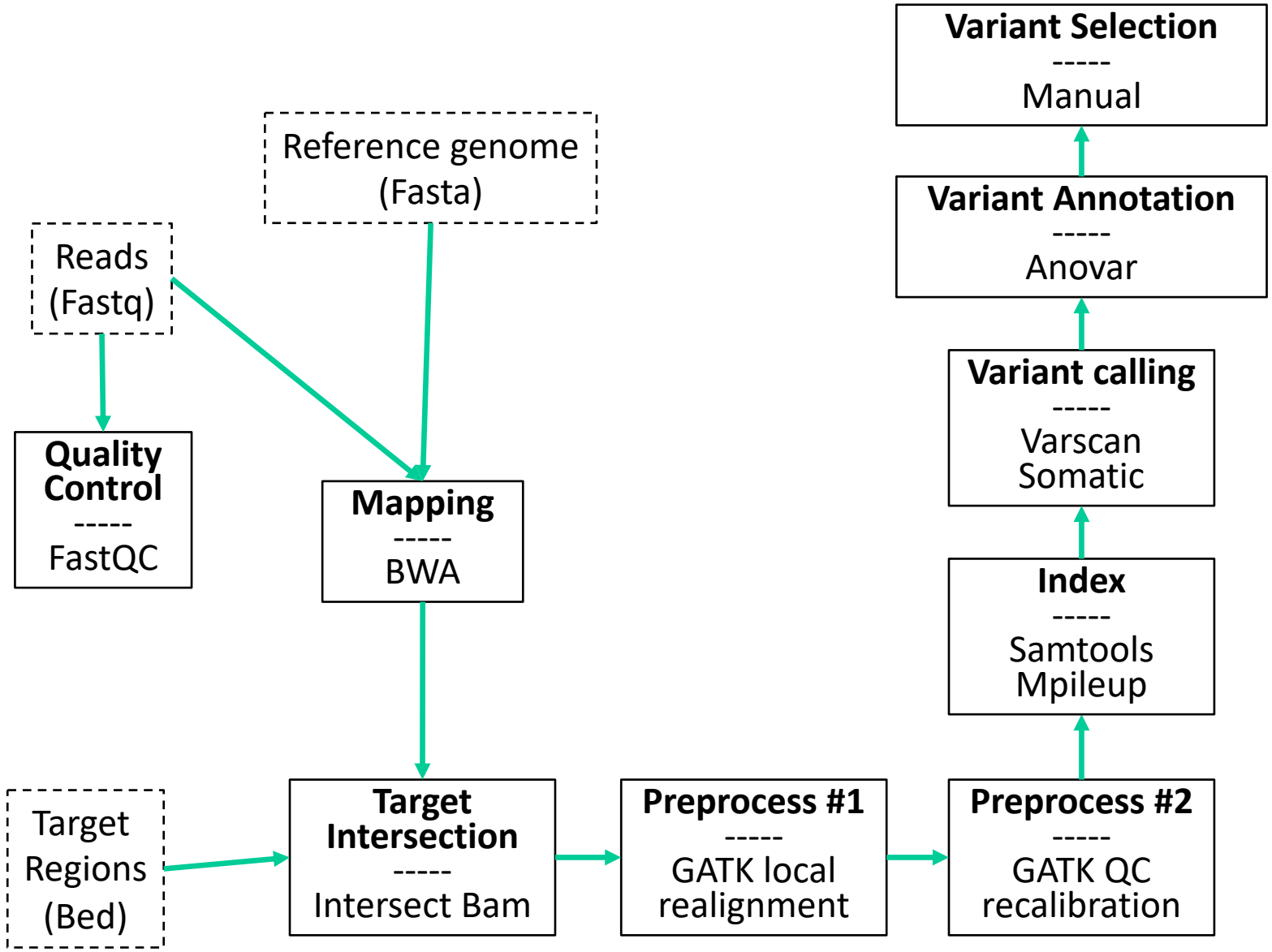
Sequencer quoi?

- Panel de gènes
  - Une série d'exons d'intérêt (gènes de cancer= 100kb)
- Exome
  - Tous les exons du génome (30 Mb)
- Whole genome
  - Le génome complet (3 Gb)

Un pipeline « Variants »

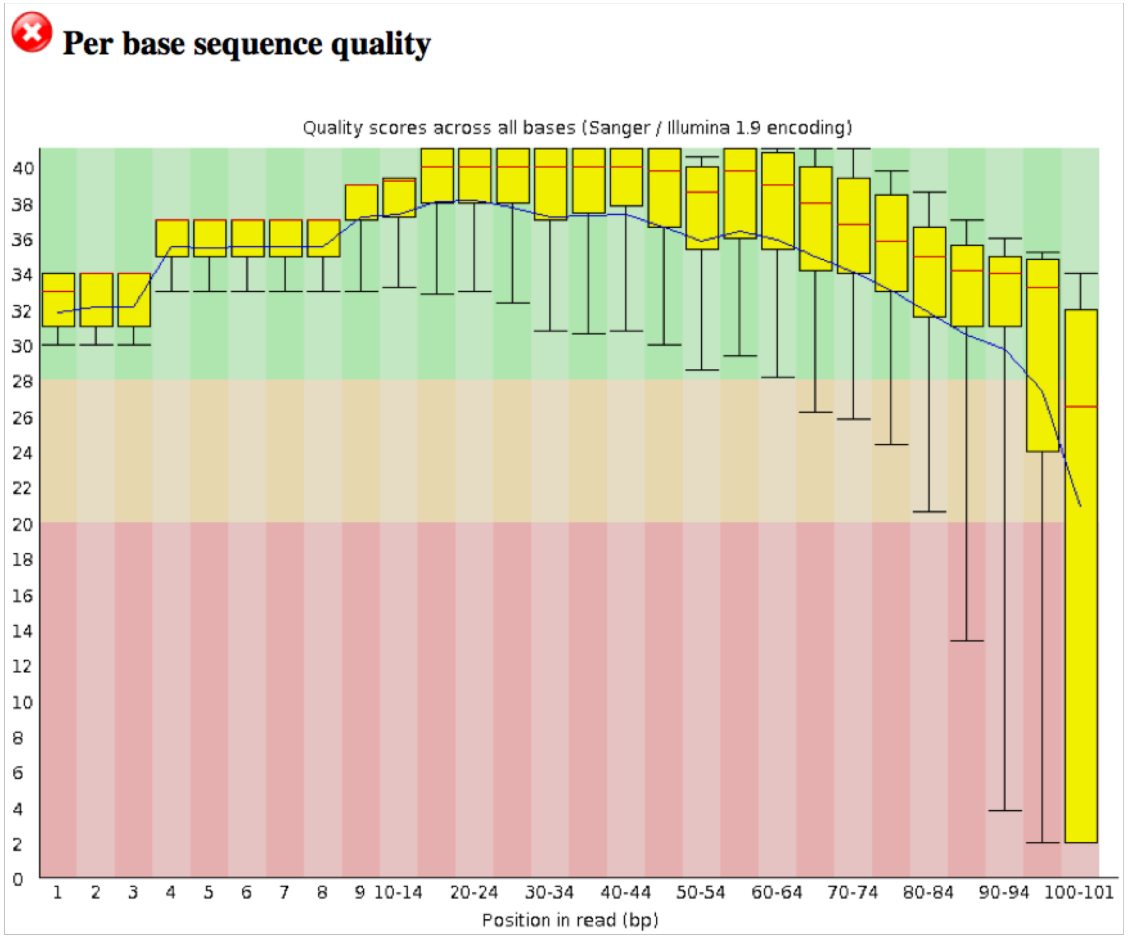


Un vrai pipeline « variants »



FastQC Metrics

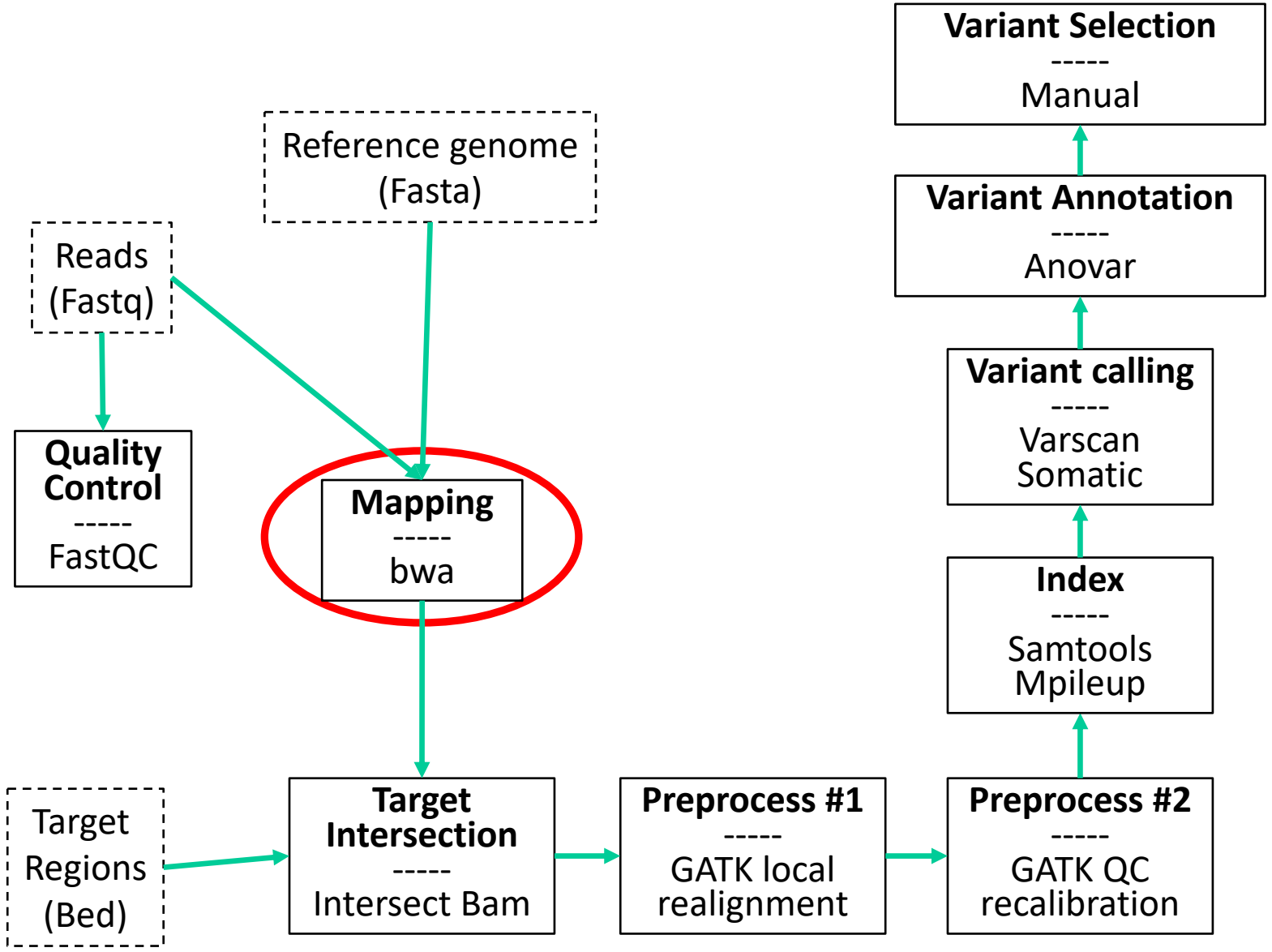
- Look at the different metrics for both reads
- **Problem**: the per base sequence quality of the Read2 are quite low towards the end



Solution:
Trim the 25bp from
the 3' end
of the reads

➢ Higher confidence in the sequenced information

(Trimmomatic)


Variant Selection

Manual
Variant Annotation

Anovar
Variant calling

Varscan
Somatic
Index

Samtools
Mpileup
Preprocess #2

GATK QC
recalibration
Preprocess #1

GATK local
realignment
Target
Intersection

Intersect Bam
Target
Regions
(Bed)
Mapping
---
bwa
Quality
Control

FastQC
Reads
(Fastq)
Reference genome
(Fasta)

Most popular aligners for variant analysis

(support mismatched, gapped, paired-end alignment)

- BWA
  - Li H. and Durbin R. (2009)
- Bowtie2
  - Langmead B, Salzberg S (2012)

Mapping with bwa

Use « bwa» to align reads on the hg19 genome

Example:

bwa mem -M -t 2 -A 2 -E 1 ~/prof/Database/hg19.fa analysis/normal_R1.fastq analysis/normal_R2-trimmed.fastq

SAM>BAM, sorting and indexing

- Use samtools for
  - making BAM files
  - Sorting
  - indexing reads

Fichier BAM

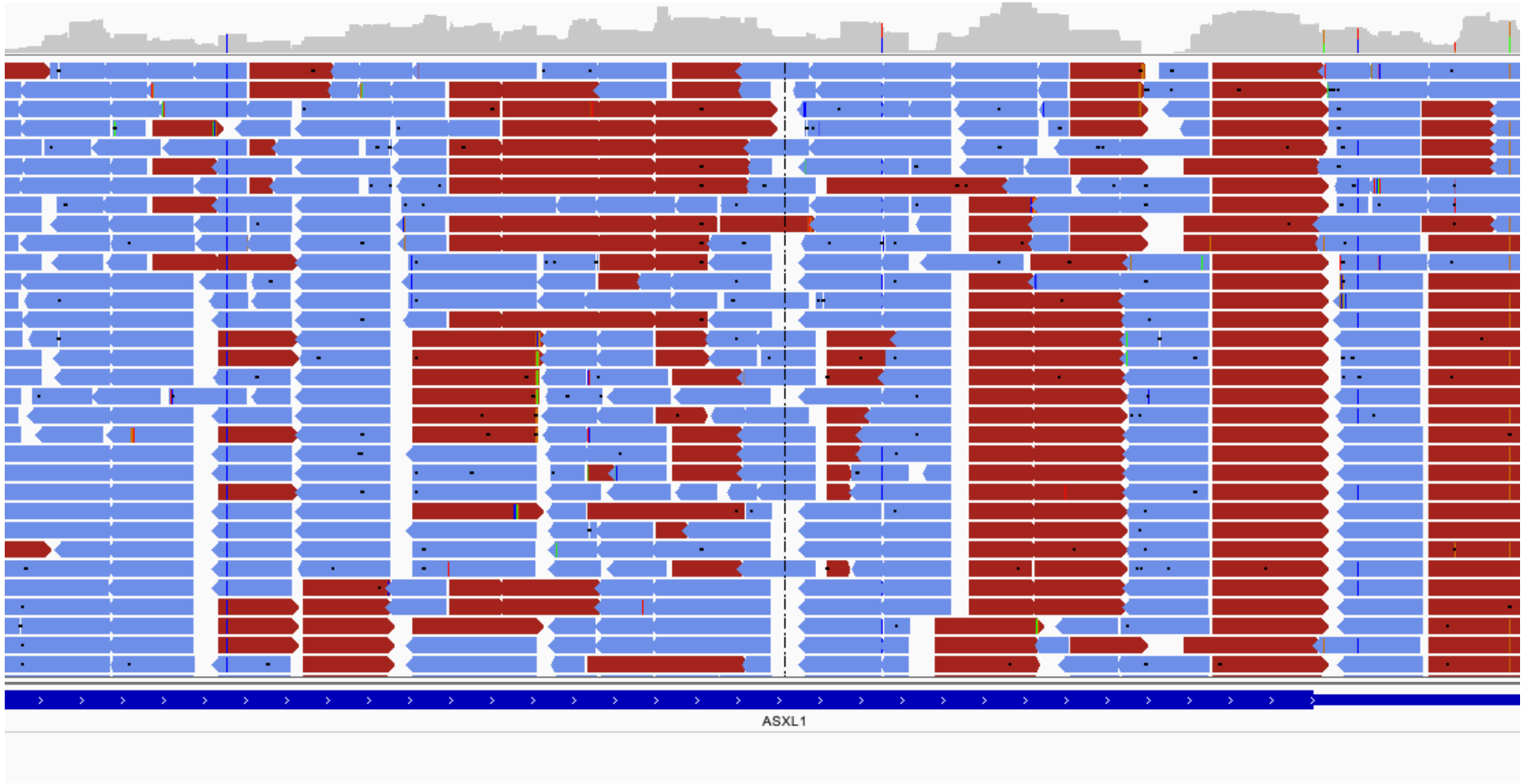
```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Mapping statistics

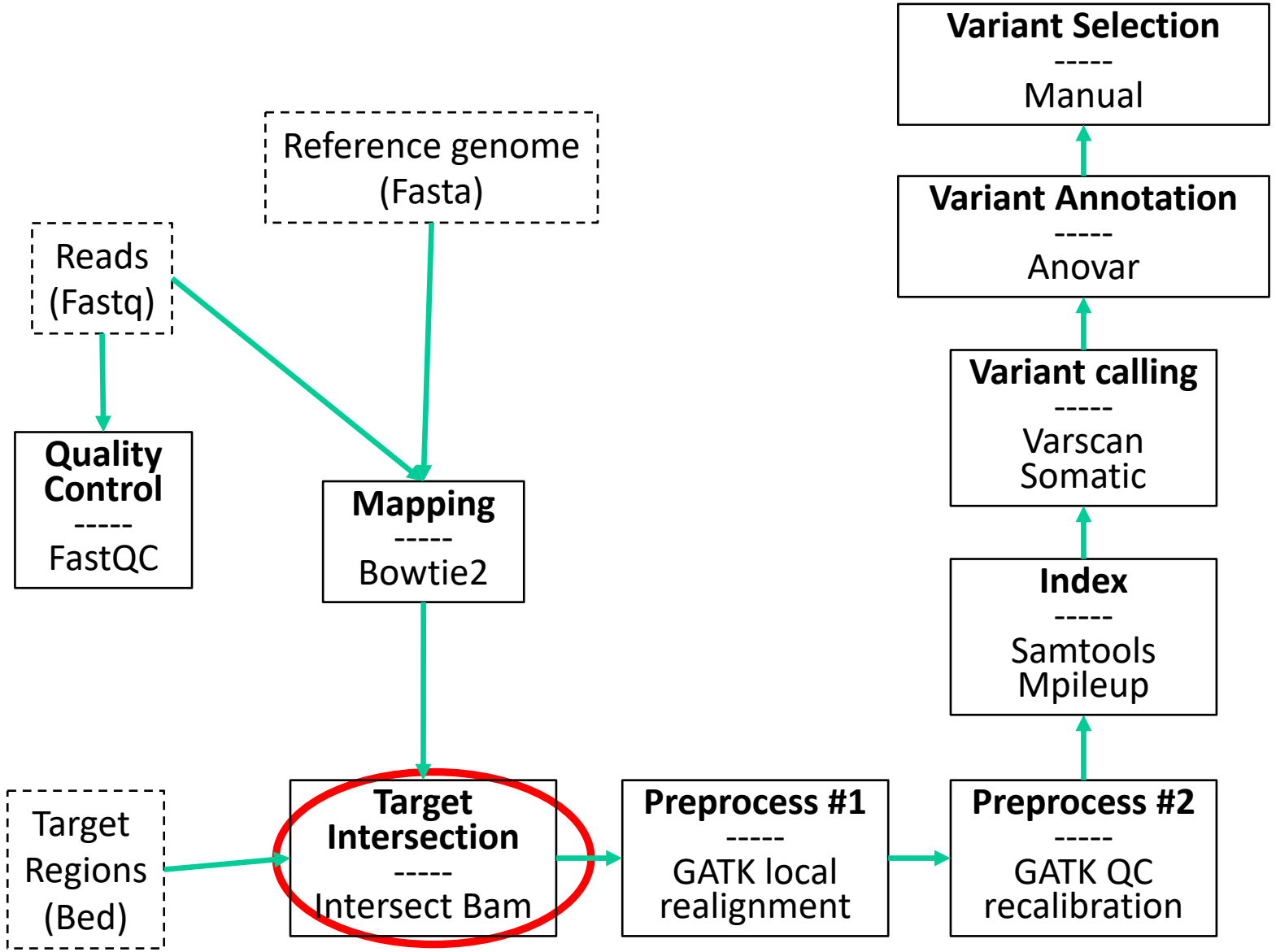
- Use « Samtools flagstat » to get mapping statistics

```
[root@vm0079 analysis]# samtools flagstat normal.sorted.bam
90662 + 0 in total (QC-passed reads + QC-failed reads)
16 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
90503 + 0 mapped (99.82%:-nan%)
90646 + 0 paired in sequencing
45323 + 0 read1
45323 + 0 read2
90120 + 0 properly paired (99.42%:-nan%)
90366 + 0 with itself and mate mapped
121 + 0 singletons (0.13%:-nan%)
12 + 0 with mate mapped to a different chr
11 + 0 with mate mapped to a different chr (mapQ>=5)
```

Reads alignés sur le génome



BAM improvement


Variant Selection

Manual
Variant Annotation

Anovar
Variant calling

Varscan
Somatic
Index

Samtools
Mpileup
Preprocess #2

GATK QC
recalibration
Preprocess #1

GATK local
realignment
Target
Intersection

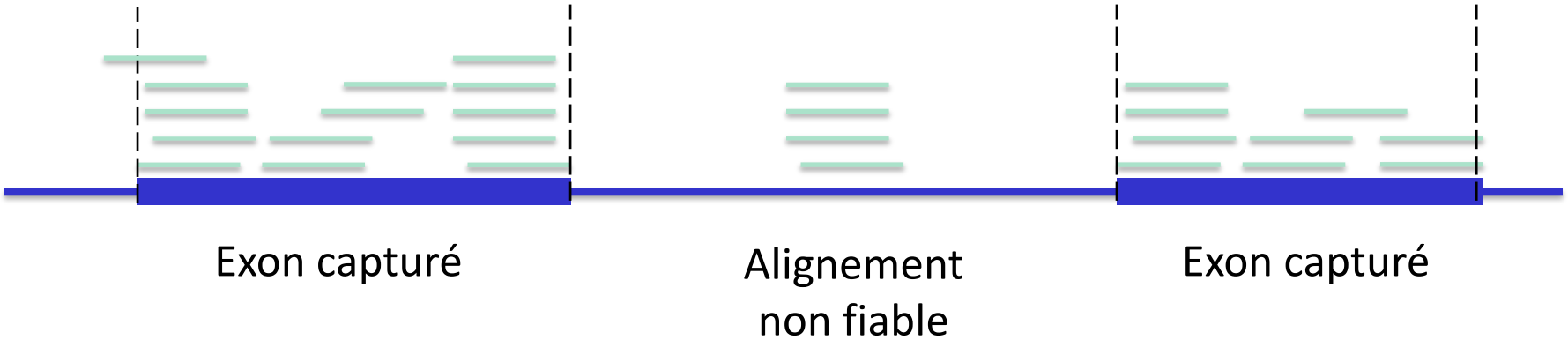
Intersect Bam
Target
Regions
(Bed)
Quality
Control

FastQC
Mapping

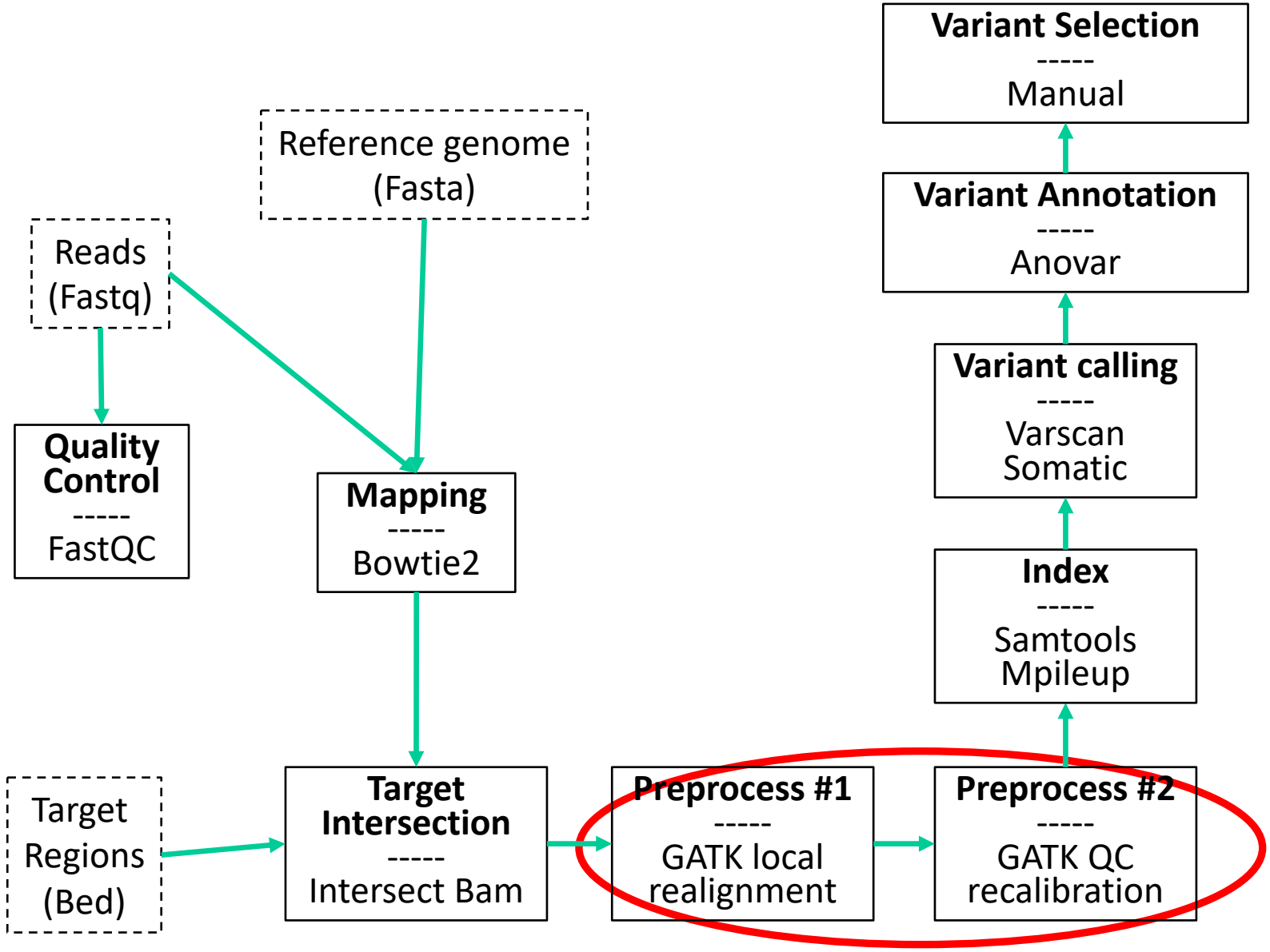
Bowtie2
Reads
(Fastq)
Reference genome
(Fasta)

Target intersection

- Comparer l'alignement obtenu à la liste des positions visées par le protocole de capture



Avec bedtools


Variant Selection

Manual
Variant Annotation

Anovar
Variant calling

Varscan
Somatic
Index

Samtools
Mpileup
Preprocess #2

GATK QC
recalibration
Preprocess #1

GATK local
realignment
Target
Intersection

Intersect Bam
Target
Regions
(Bed)
Quality
Control

FastQC
Mapping

Bowtie2
Reads
(Fastq)
Reference genome
(Fasta)

Why realigning around indels ?

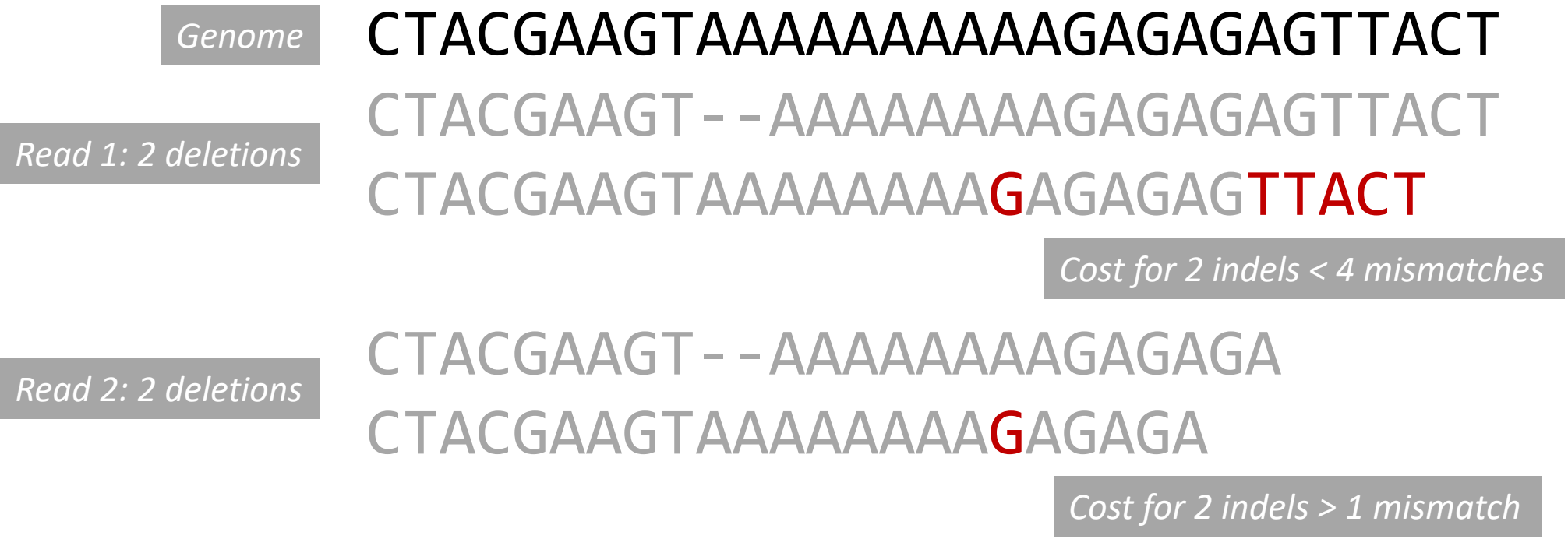
• Small Insertion/deletion (Indels) in reads (especially near the ends) can trick the mappers into wrong alignments

- Alignment scoring – cheaper to introduce multiple Single Nucleotide Variants (SNVs) than an indel: induce a lot of false positive SNVs

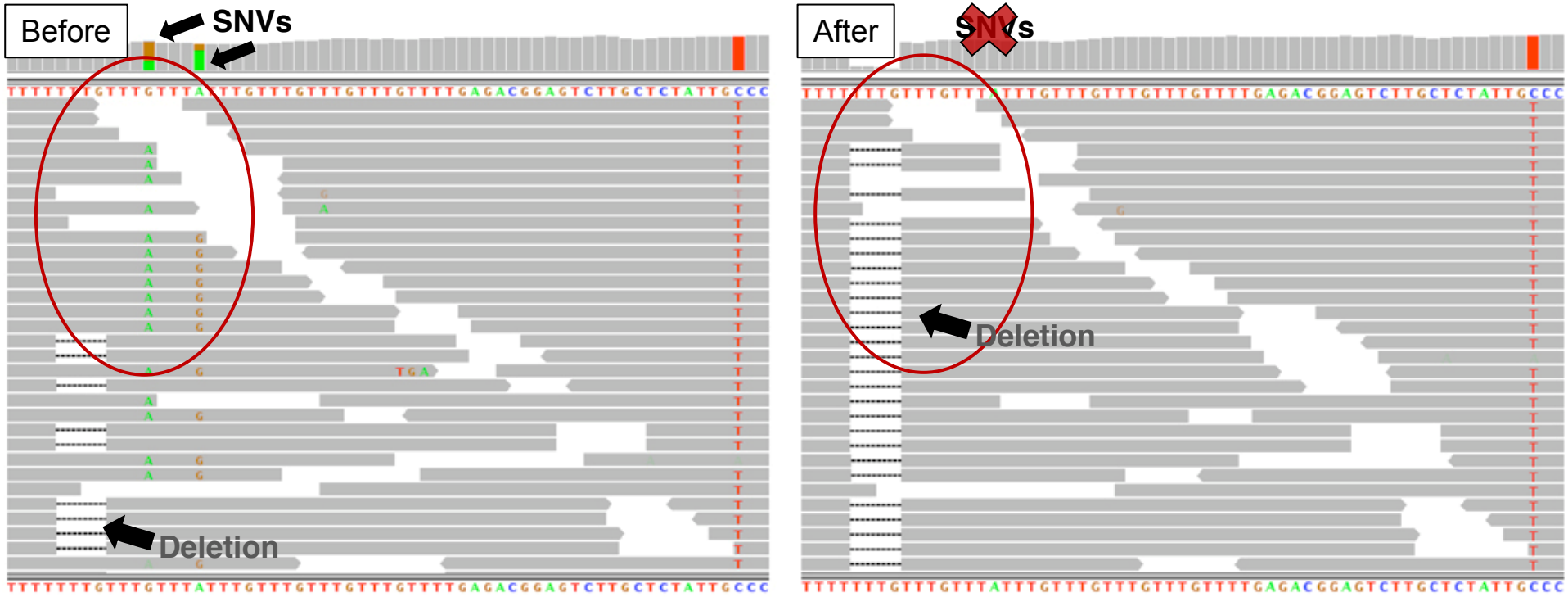
➔ artifactual mismatches

• **Realignment around indels helps improve the downstream processing steps**

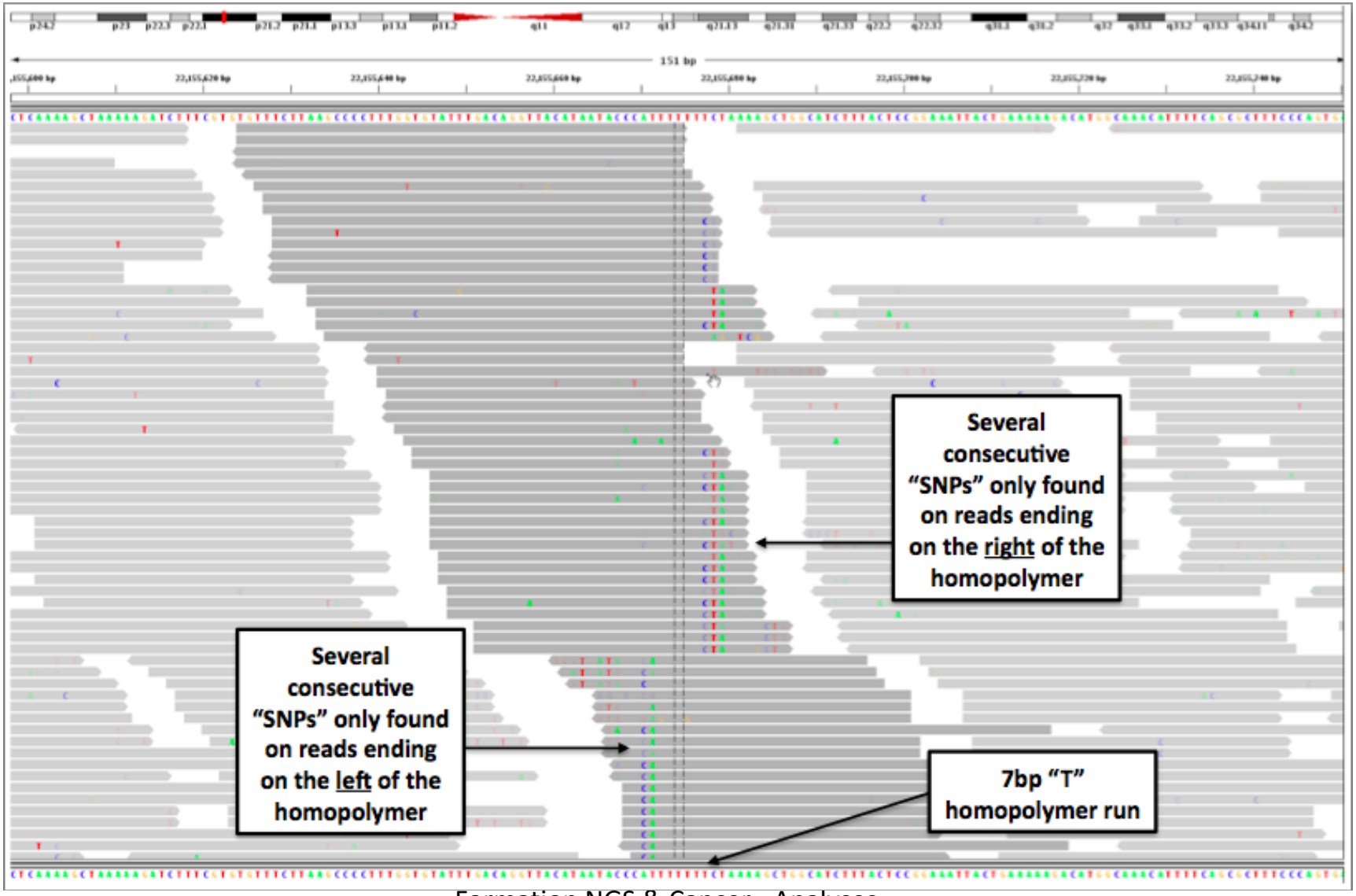
Wrong alignment near indels



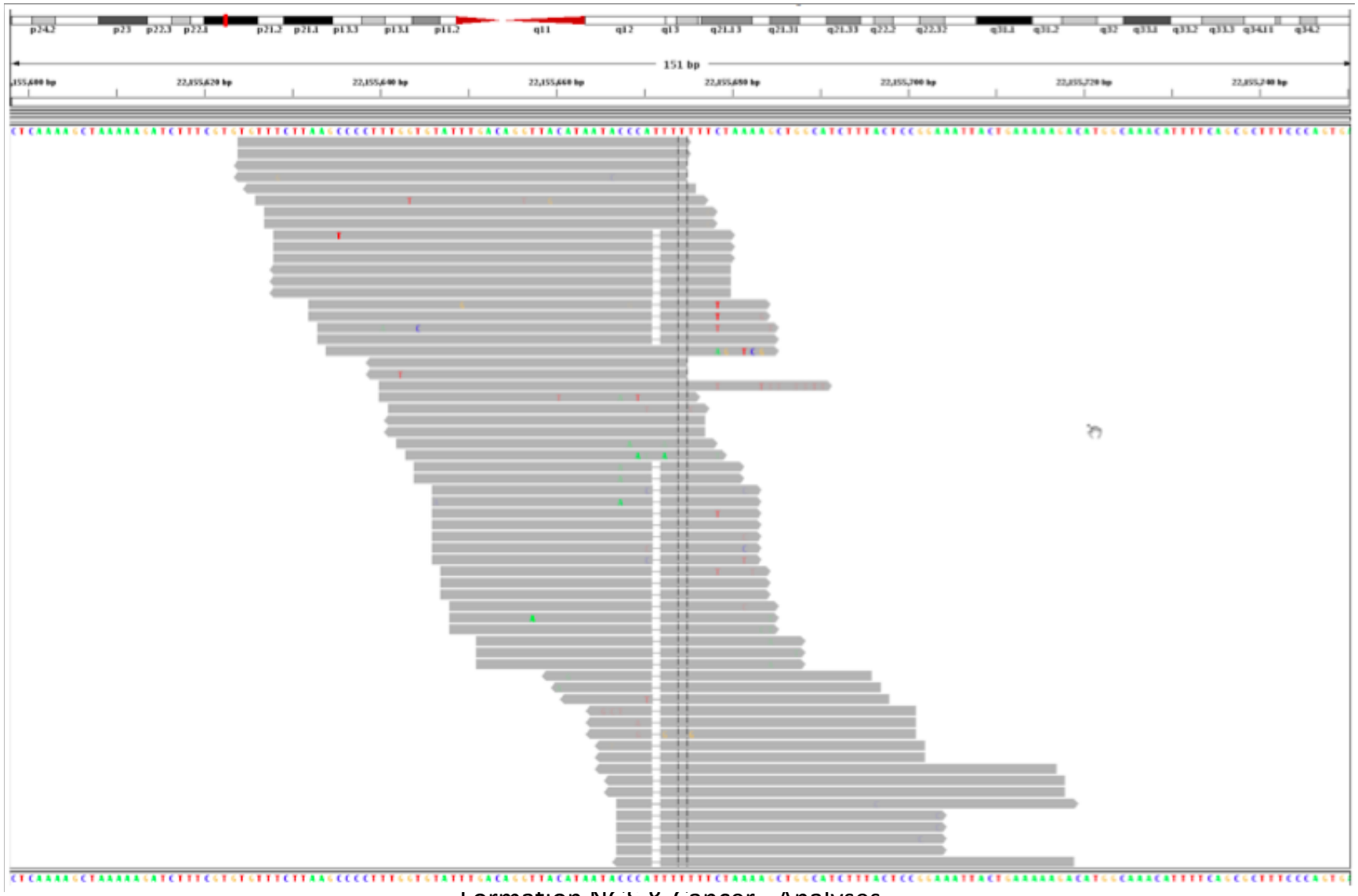
Local realignment around indels



Local realignment around indels



Local realignment around indels

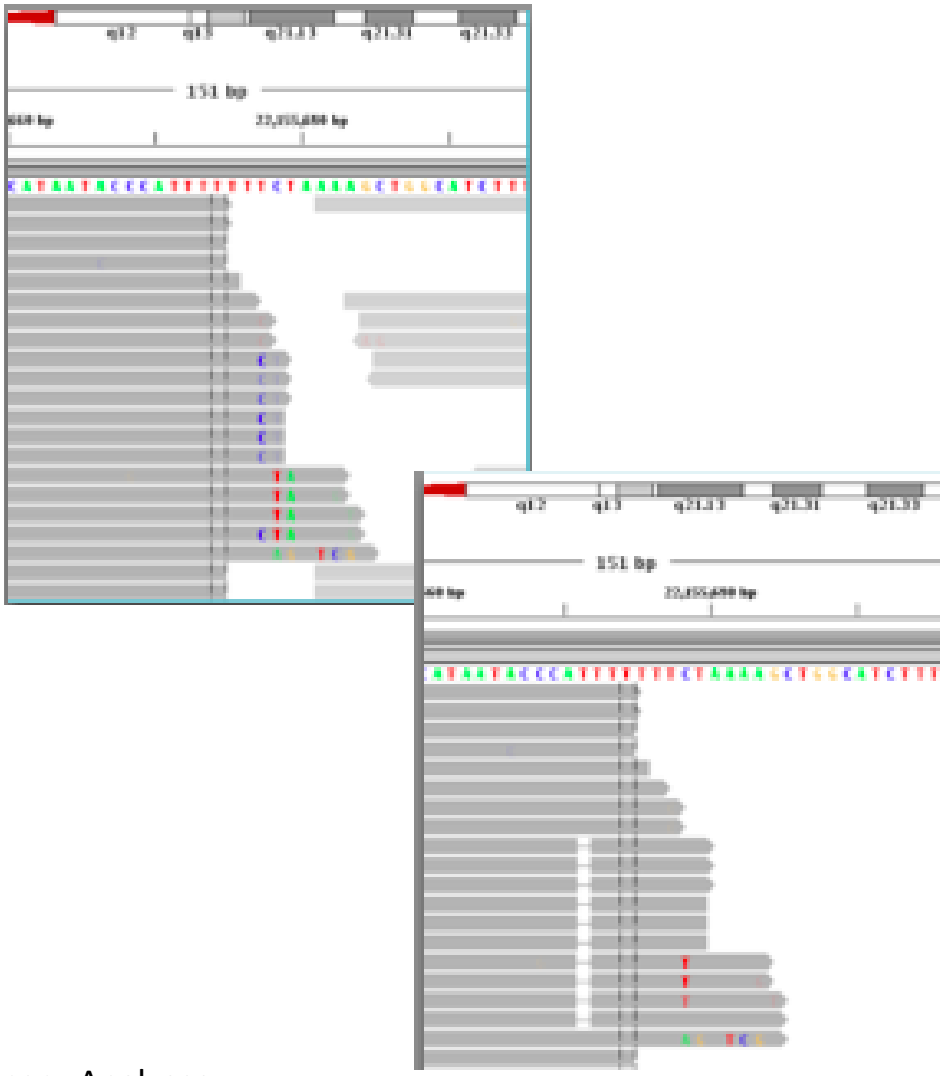


Indel realignment in 2 steps

1. Identify what regions need to be realigned
   - GATK RealignerTargetCreator
     + known sites

Intervals

2. Perform the actual realignment (BAM output)
   - GATK IndelRealigner

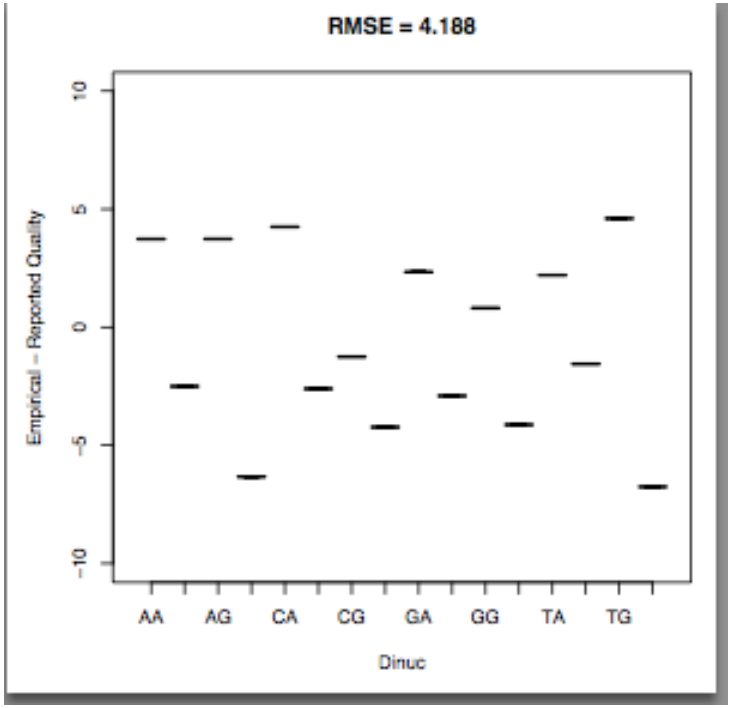


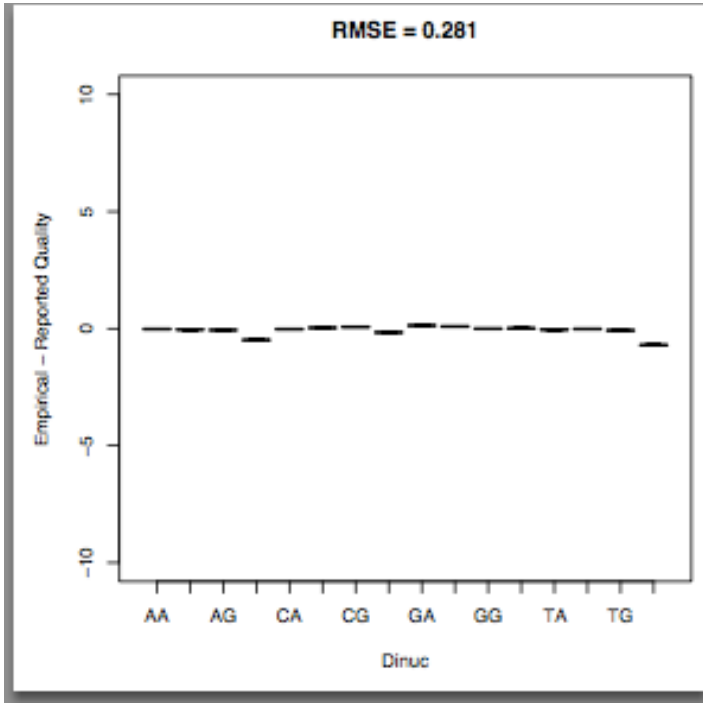
Types of realignment targets

- Indels seen in original alignments (in CIGAR, indicated by I for Insertion or D for Deletion)
- Sites where evidences suggest a hidden indel (SNV abundance)
- Known sites:
  - Common polymorphisms: dbSNP, 1000Genomes

Quality scores issued by sequencers are biased

- Quality scores are critical for all downstream analysis
- Systematic biases are a major contributor to bad calls
- Example of sequence context bias in the reported qualities:





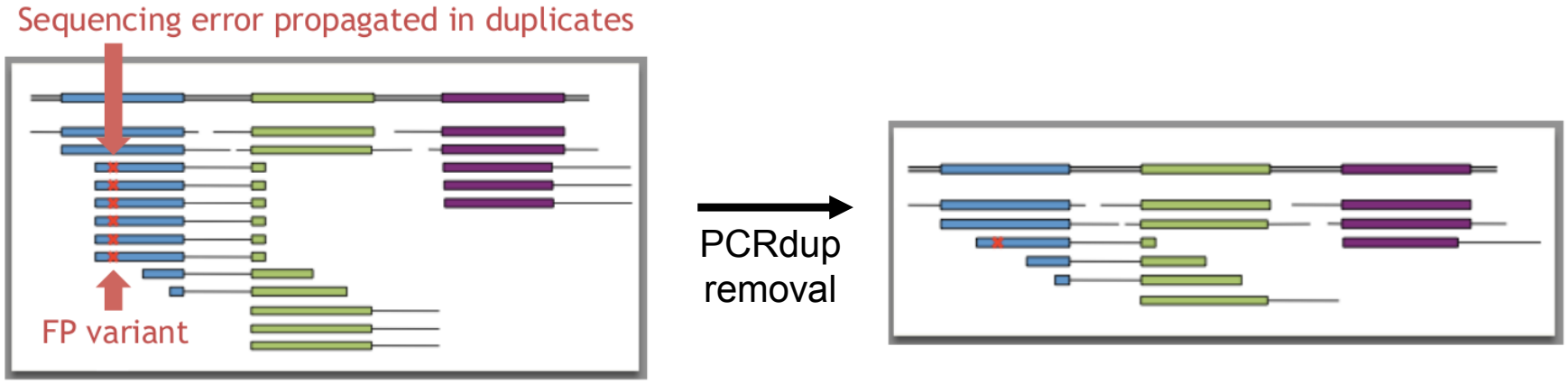
before

after

➢ GATK baserecalibrator

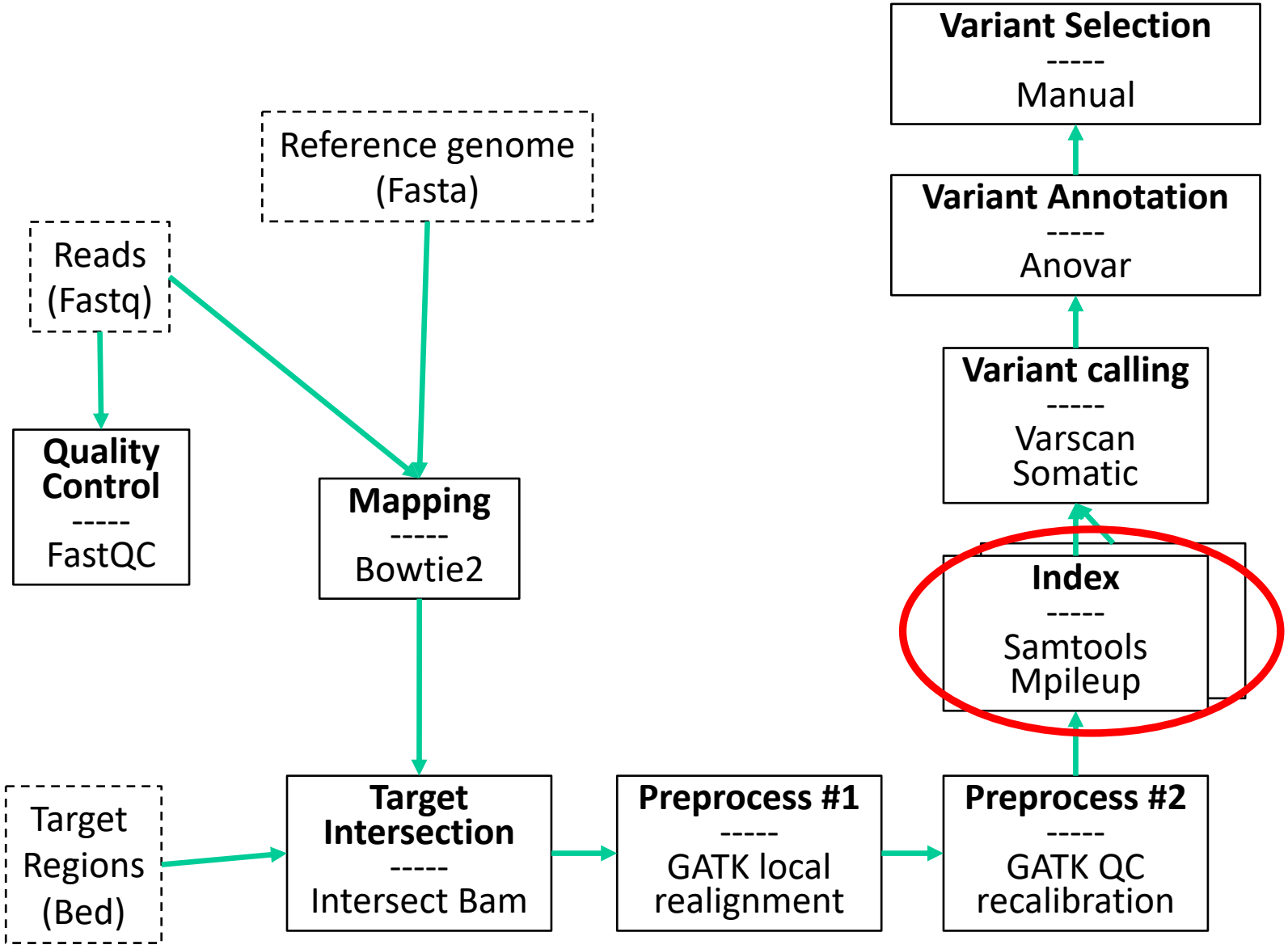
Removing duplicates (not for targeted sequencing)

- **Duplicates reads**: different reads having the same sequence caused by PCR amplification during sequencing library preparation
- The removal of the duplicates depends on the application (not suitable for sequencing on small target)



- Use "samtools rmdup" to **remove** duplicates
- Or run "samtools Flagstat" to see the number of PCR duplicates

```
9062 + 0 duplicates
```


Variant Selection

Manual
Variant Annotation

Anovar
Variant calling

Varscan
Somatic
Index

Samtools
Mpileup
Preprocess #2

GATK QC
recalibration
Preprocess #1

GATK local
realignment
Target
Intersection

Intersect Bam
Target
Regions
(Bed)
Quality
Control

FastQC
Mapping

Bowtie2
Reads
(Fastq)
Reference genome
(Fasta)

Pileup format

Describes base-pair information at each position

Reference base

Base qualities

chr12	112888238	A	108	.$,$.,,.,,..,,.,.,..,.,,.,.,,.,..,,..,,.,.,.,.,,.,.,.,,,,.,. ,.,.,,,.,.,.,.,.,,.,.,.,.,,,,.,,.,.,..,,,,,,.,.,,,,^F,	=4 =??????@???@?@@ @=@ ??@ ? @? ? ??< ? ??@??????? ? @??? ? ??@?? A???@@ ?@@???AB????= ? @ @@??@@?@ A 00
chr12	112888239	C	108	.$t,.,,.T,tT,.,T.,.,t.tTtt.tTT,t.T,tTt.tT,T,,.t TtTttt.,.,Tt.ttt.,T,.,.tT,,T,T,.tT,,t,TttTtT,T. ,tt,,T,T,tttt^F.^F,	936 78??6??6 45<875? ??? ?@6 @6???<?=6 66= ? ???? 6=7??=???<8@7=7=? 77?7?8 ??78?7????? 8 <8??88 9?8 ?0048

Number of reads covering the site (total depth)

Read bases:

. / , = match on forward/reverse strand

ACGTN / acgtn = mismatch on forward/reverse strand

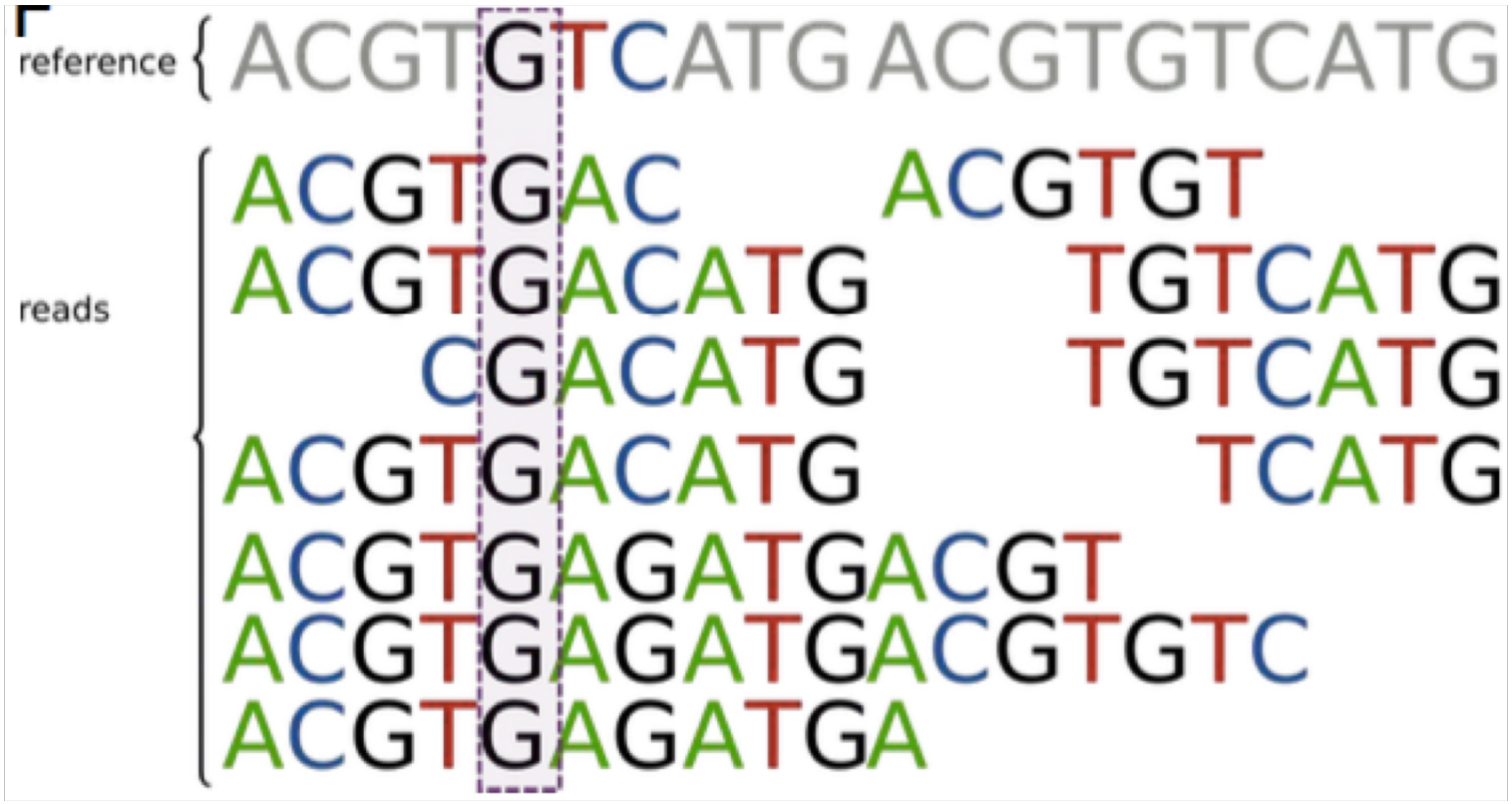
`-\+[0-9]+[ACGTNacgtn]+' indicates an indel

daniel — root@vm0079:~/mydisk/AtelierNGS/Variants/analysis — ssh -A -p 22 root@192.54.201.128 — 148×46

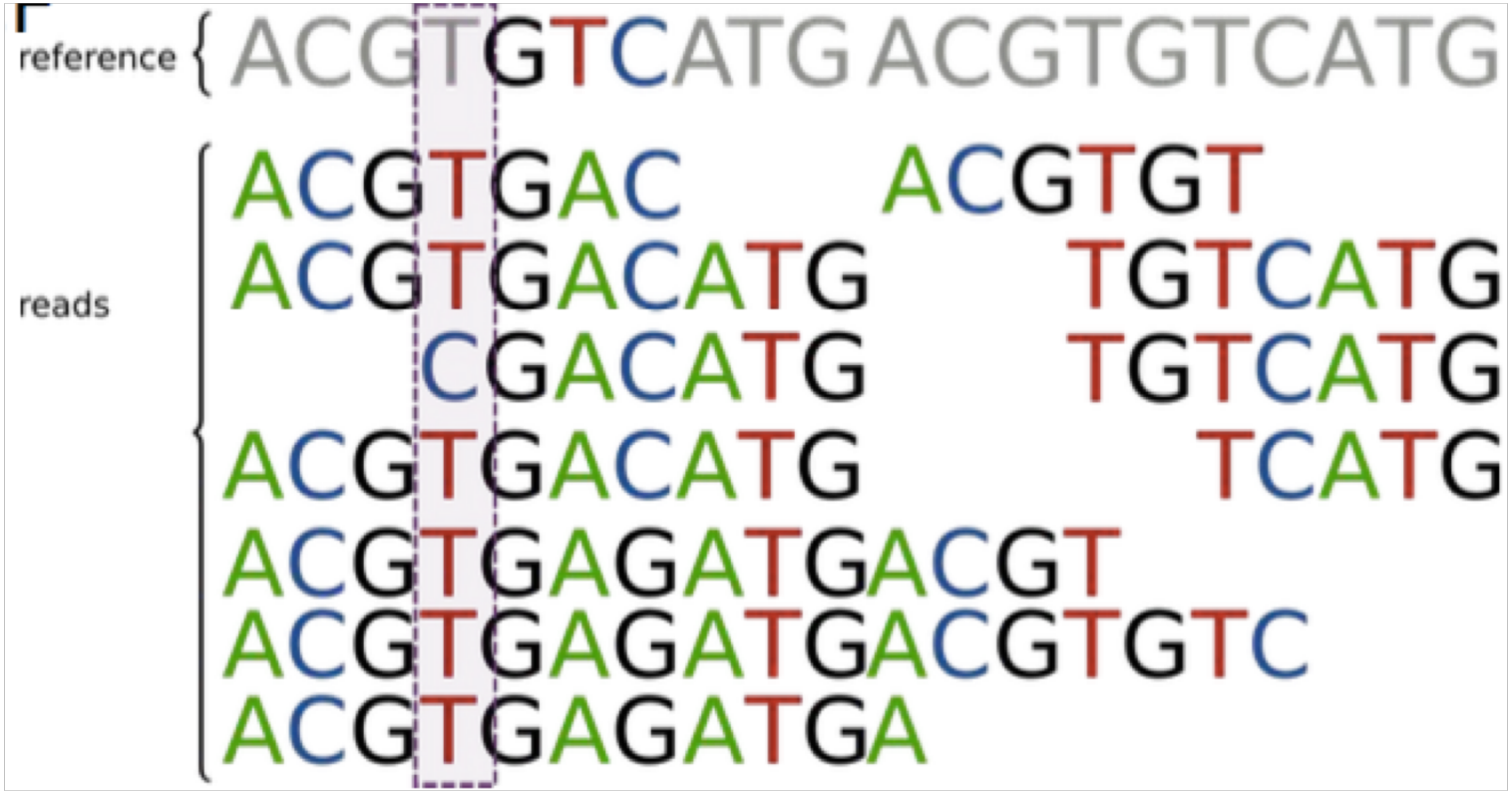
root@vm0079:~/mydisk/AtelierNGS/Variants/analysis — ssh -A -p 22 root@192.54.201... | ~ — -bash

```
File Edit Options Buffers Tools Help
chr11   26692809    T   65   ,$,$,$.,,.,,,,,,,,,,,,,.,,,,.,.,,,,,,,,..,.,,,,.,,,,,,,,,,,,,,,.,,,,,,,   @C@cFFCFHHGHIG9JIJJDA?Ge@fJHGJmIFIj$
chr11   26692810    A   62   .$,,.,,,,,,,,,,,,,.,,,,.,,,,,,,,.,.,,,,,,.,,,,,,,,,,,,,,,,.,,,,,,,,^].   hFFCHGHHHFG<JIJGDCHJdEeJJFJkIFHljJq$
chr11   26692811    C   62   ,gGgg,gg,gggggg,G,g,G,.g,gg,g.,..gGgg.ggg,,g,ggg,,,,g.,,,,gg,G   CC>HHHHHHG?JIIIJDCGI`@eJJIIlJEElkIpI?[J:IBJ$
chr11   26692812    A   64   ,,.,,,,,,,,,,,,,,.,,,,.,,,,,,,,..,.,,,,,,.,,,,,,,,,,,,,,,.,,,,,,,,.   CC@DGHHHGD<JJJIJ@CGJaIgJJIJlJDIjiJq$
chr11   26692813    A   62   ,$,$.,,,,,,,,,,,,,.,,,,.,,,,,,,,..,.,,,,,,.,,,,,,,,,,,,,,.,,,,,,,,.   CCCDFHHHHH>HHHGJAFEJbEfIJIJkJEHm_Io$
chr11   26692814    A   62   .,,,,,,,,,,,,,,,,,.,,,,.,,,,,,,,..,.,,,,,,.,,,,,,,,,,,,,,.,,,,,,,.^].   CDFFFFHGCHHHHHCCHIUBfJJIJnJEDkfInHF$
chr11   26692815    G   61   .$,,,,,,,,,,,,,,,,.,,,,.,,,,,,,,..,.,,,,,,.,,,,,,,,,,,,,,.,,,,,,,,..   C=FFFFFH:HHHHHC?EHOGeJIIJkICCmhGnEBjII1GBII$
chr11   26692816    A   62   ,,,,,,,,,,,,,,,,,,.,,,,.,,,,,,,,..,.,,,,,,.,,,,,,,,,,,,,,.,,,,,,,,..   BDFFFDH2HFHHHDADHZBgIJIHkGEFmgHmAFChIDG1DBI$
chr11   26692817    C   62   ,,,,,,,,,,,,,,,,,,.,,,,.,,,,,,,,..,.,,,,,,.,,,,,,,,,,,,,,.,,,,,,,,..   ?FFFFDD=FHHDHDC<H^GfJJJJjIDCkgJoCA2fIDI9IEH$
chr11   26692818    A   60   ,,,,,,,,,,,,,,,,,,.,,,,.,,,,,,,,..,.,,,,,,.,,,,,,,,,,,,,,.,,,,,,..^],   ?BDDFDFD?DAFC=?H]EaHJHIiHDEjfHmAF[I?GHIJHJ9$
chr11   26692819    C   63   ,$,,,,,,,,,,,,,,,,.,,,,.,,,,,,,,..,.,,,,,,.,,,,,,,,,,,,,,.,,,,,,,,,.,   ?@CCCFDBFFFEFDDFHiHgJJIJmJDJmjJn999$
chr11   26692820    A   62   ,$,,,,,,,,,,,,,,,,.,,,,.,,,,,,,,..,.,,,,,,.,,,,,,,,,,,,,,.,,,,,,,,,.,   @C@CCD?FFFDFD?BHiGiIJGJmJDIliJnDH<kJFH:IDJJ$
chr11   26692821    T   61   ,$,$,$,,,,,,,,,,,,.,,,,.,,,,,,,,..,.,,,,,,.,,,,,,,,,,,,,,.,,,,,,,,,.,   C@CCFDFFFBFCADFhHfHJIJjICHlhInEE2iJ$
chr11   26692822    A   58   ,$,,,,,,,,,,,,,,,,.,,,,.,,,,,,,,..,.,,,,,,.,,,,,,,,,,,,,,.,,,,,.,   BC?FFF:FCDDFhHfHJJIjJCClhJn9A<jJEJEI:JJJFII$
chr11   26692823    T   57   ,,,,,,,,,,,,,,,,,,.,,,,.,,,,,,,,..,.,,,,,,.,,,,,,,,,,,,,,.,,,,.,   C;CCCFFEDDFcFfHJJJkJDHnjJnCH<gJEJFJ<IJGEHHJ$
chr11   26692824    T   57   ,$,$,,,,,,,,,,,,,,.,,,,.,,,,,,,,..,.,,,,,,.,,,,,,,,,,,,,,.,,,,,.,   C8C@C@CCBDDg8hHHIJkIEBmjJm9E<fJ>G?JDHIJIIFI$
chr11   26692825    A   55   ,$,$,$,,,,,,,,,,,,.,,,,.,,,,,,,,..,.,,,,,,.,,,,,,,,,,,,,.,,,,.,   B@@@CC@DFdFjHHHJkJFFmlJoHGHiJDJ@JGIJJHJ@JJJ$
chr11   26692826    G   52   ,$,$.,,,,.,,.,,,,,,,,.,,..,,,,,,.,,,,,,,,,,,,,,.,,,,,,..,   @BA?@B]DiFHDHjI@GnlJnHF@fJCJ@JFHJJGJBIJJJHH?F:EEFjH$
chr11   26692827    C   50   .$,$,,.,,,,,.,,,.,,,,,,,,.,,,,,,,,,,,,,,,.,,,,,..,   9??@UBiFHFHiHCEmjJgIBAeI<JBJADJJHJGJJJJHHEC>FEFcJC
chr11   26692828    A   48   ,$,$.,,.,,,,,.,,.,,,,,,,,,,,,,,,,,,,,,.,,,,,,..,   ?@^DiFHGHiHDImjJlH@<fJ9JHJDFIJEJIJJJJHHEE;FCFnJ@
chr11   26692829    A   46   .,.,,,,,,,,,.,,,.,,,,,,,,.,,,,,,,,,,,,,,,,,,..,   g@iFFHHiHDHlkJmCFIiJ<JFJAICJIJFJIJJHEEH;DDFnJ>
chr11   26692830    G   46   .,.,,,,,,,,..,,,.,,,,,,,,.,,,,,,,,,,,,,,,,,,..,   G;hFFFHiHDDkkJl4HAcJ3JBJHIHJIJIJJJJGFEF?ADEmJC
chr11   26692831    A   46   .$,$.,,,,,,,.,,.,,,,,,,,,.,,,,,,,,,,,,,,,,,..,   I?kCFDFfGDHlmJmFG9hJFJ@JDIEJHIIJJJJGHFHEFAEmJE
chr11   26692832    T   46   .,,,,,,,.,..,,,,.,,,,,,,,.,,,,,,,,,,,,,,,,..,^],   eCFDFfFDHmmJj<FIgJAJ@JFIEJEIGIJJJJJEHHAECGmJED
chr11   26692833    A   46   .$,$,,,,$,.,..,,,,.,,,,,,,,,,,,,,,,,,,,,,,..,^],   hCFDFfFDGmmJkCEHiJJEJBFGJHICIJJJIJIGC=E=HmJCDD
chr11   26692834    T   43   ,,,,,.,,.,,,.,,,,,,,,,,,,,,,,,.,,,,,,,,..,,,   CDFFDFmnHm3AGiJ9I:IBIGJEJDJJJJJHGG=HHHmJEBD
chr11   26692835    C   45   ,,,,,.,,.,,,.,,,,,,,,,,,,,,,,,.,,,,,,,,,..,,,   CCFF@DmmGlFE<gJ>ICJ>JFJJJHJIJIJGGIE3EHGHmJCAD
chr11   26692836    C   45   ,$,,,.,,.,,,.,,,,,,,,,,,,,,,,,.,,,,,,,,,..,,,   CBCFDBnnHk<CEbJCJ?IEJIJFIGJGJJJDIGD@CCIHnJ@CD
chr11   26692837    T   44   ,$,,.,.,,,,.,,,,,,,,,,,,,,,,,.,,,,,,,,,..,,,   @CCDFmmHmD>DhJ<I:FBJJJIJFJJJJJ>GE=:I@IJmJCCC
chr11   26692838    T   42   ,$,.,..,.,,,,.,,,,,,,,,,,,,,,.,,,,,,,,..,,,   CCDFmmHm4FAeJ:H?JGJIJIIIIJIJJCIGGHDJInJFDC
chr11   26692839    G   42   ,$.,..,.,,,,.,,,,,,,,,,,,,,,,.,,,,,,,,..,,,   CD@llFmF?AkJFJ?JHJJJIJIJJJJJGGICEEGJFnJFDD
chr11   26692840    A   41   .$,..,.,,,,.,,,,,,,,,,,,,,,,.,,,,,,,,..,,,   CBmlFm>A?kJBJ@IJJJJGJIJJJCIGCHAAH@JImJDEE
chr11   26692841    T   40   ,$..,.,,,,,,,,,,,,,,,,,,,,,.,,,,,,,,..,,,   @mlFmDFCkJCGIGHIIIGJGJJJHGDCHFCEGJIlJ@DD
chr11   26692842    G   39   .$.,.,,.,,,,,,,,,,,,,,,,,.,,,,,,,,..,,,   lkFjAD=jI2IAGFHGIHIFJIHIJGIF@=FIIGnJADD
chr11   26692843    T   39   .,.,,,.,,,,,,,,,,,,,,,,,.,,,,,,,,,..,,,   lDkDD=hH2HAEHAFGHCHFIHFIICGF7@FHGFmI=CC
chr11   26692844    A   36   .,.,,,.,,,,,,,,,,,,,,,,.,,,,,,,..,,,   j@iDBdFG2FFBEEGECIGDII=FG7.BFICkIAAC
chr11   26692845    C   39   .,.,,,.,,,,,,,,,,,,,,,,,.,,,,,,,,,..,,,   mC]==4iH2J3JJEJGJGHHJJJJJHHJHBCCJJoJHEE
chr11   26692846    T   38   .$,$.,,.,,,,,,,,,,,,,,,.,,,,,,,,,..,,,   kBi7DdH:G<IIGIIGGEGJJIIJGGJHHEAJInJHCD
chr11   26692847    C   38   .$,$,,.,,,,,,,,,,,,,,,.,,,,,,,,,,..,,,   j@;8fHDH<JJIJJJJJDJJJIJIE@HIICAIJoJJCE
chr11   26692848    C   38   ,,.,,,,,,,,,,,,,,,,,,.,,,,,,,,,..,,,^],   81dHDHDJJ@IJJHJJEJIJHJJJIIHHGGIJqJJEEC
chr11   26692849    A   39   ,$,$.,,.,,,,,,,,,,,,,,,.,,,,,,,,,,..,,,   ?1fFBHFGJFJJJHJJHJJJDJJIHIHFDGJIqIHCEEC
chr11   26692850    A   37   .,,,,,,,,,,,,,,,,,,,,.,,,,,,,,,,..,,,   jFDHFHH@IIJFJICJJJJJJGIICBCCJJqJG7FEA
chr11   26692851    A   37   .,,,,,,,,,,,,,,,,,,,,.,,,,,,,,,,..,,,   eFBFCHH?IJJEIJCJIJIJJGHID<C=JIqJG=FFD
-UU-:----F1  tumor.mpileup    1% L??    (Fundamental)------------------------------------------------------------
```

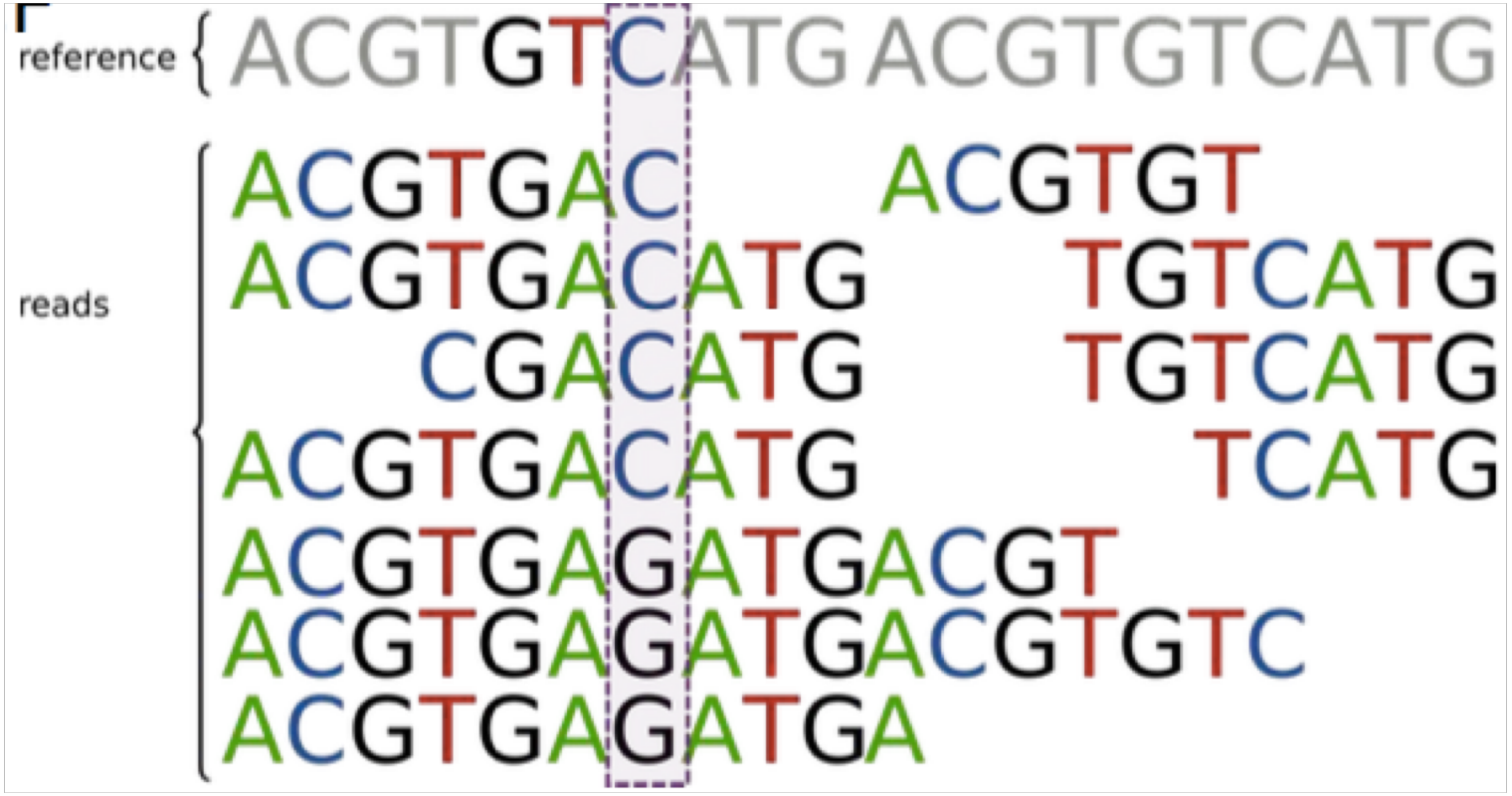
Analyse du PileUp



Analyse du PileUp

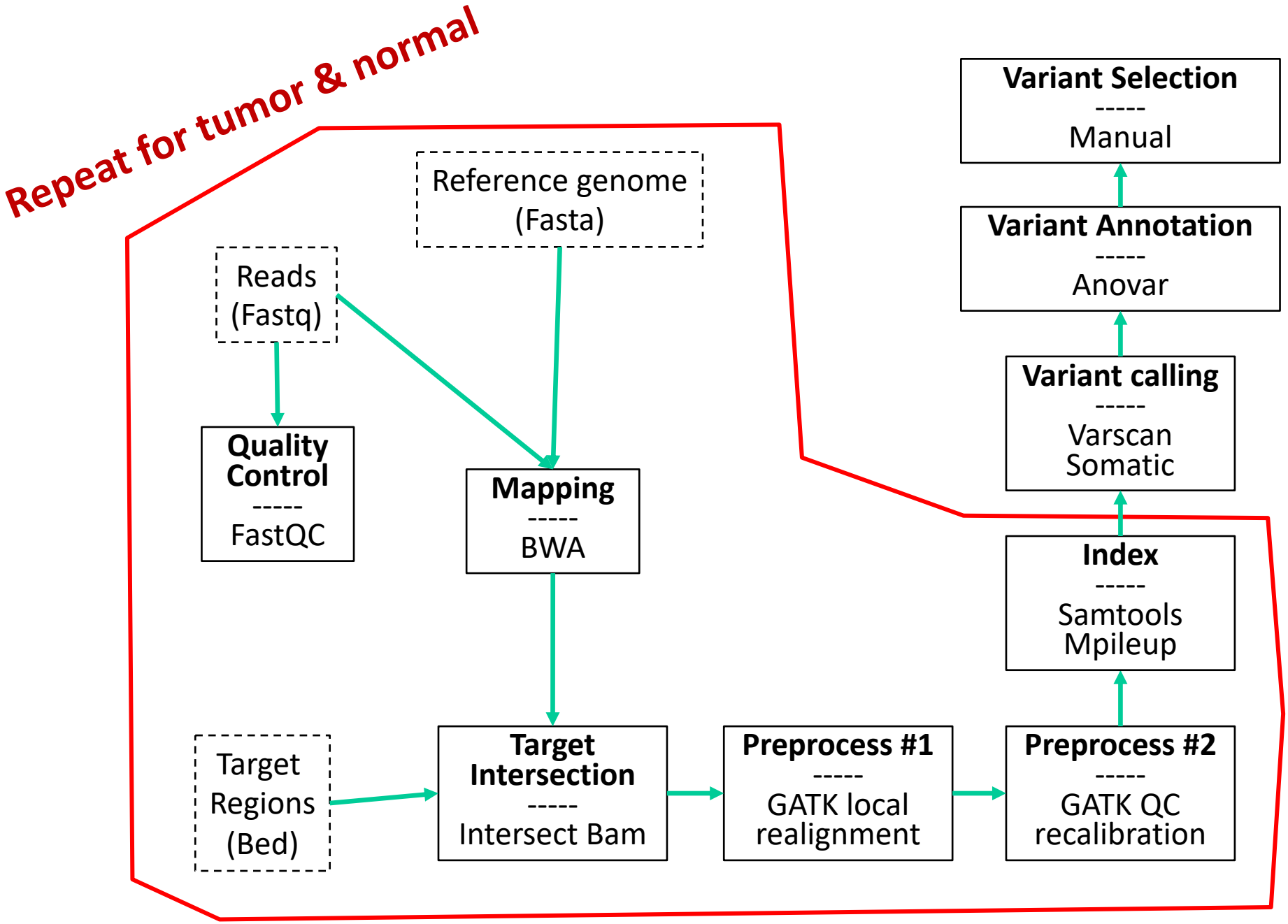


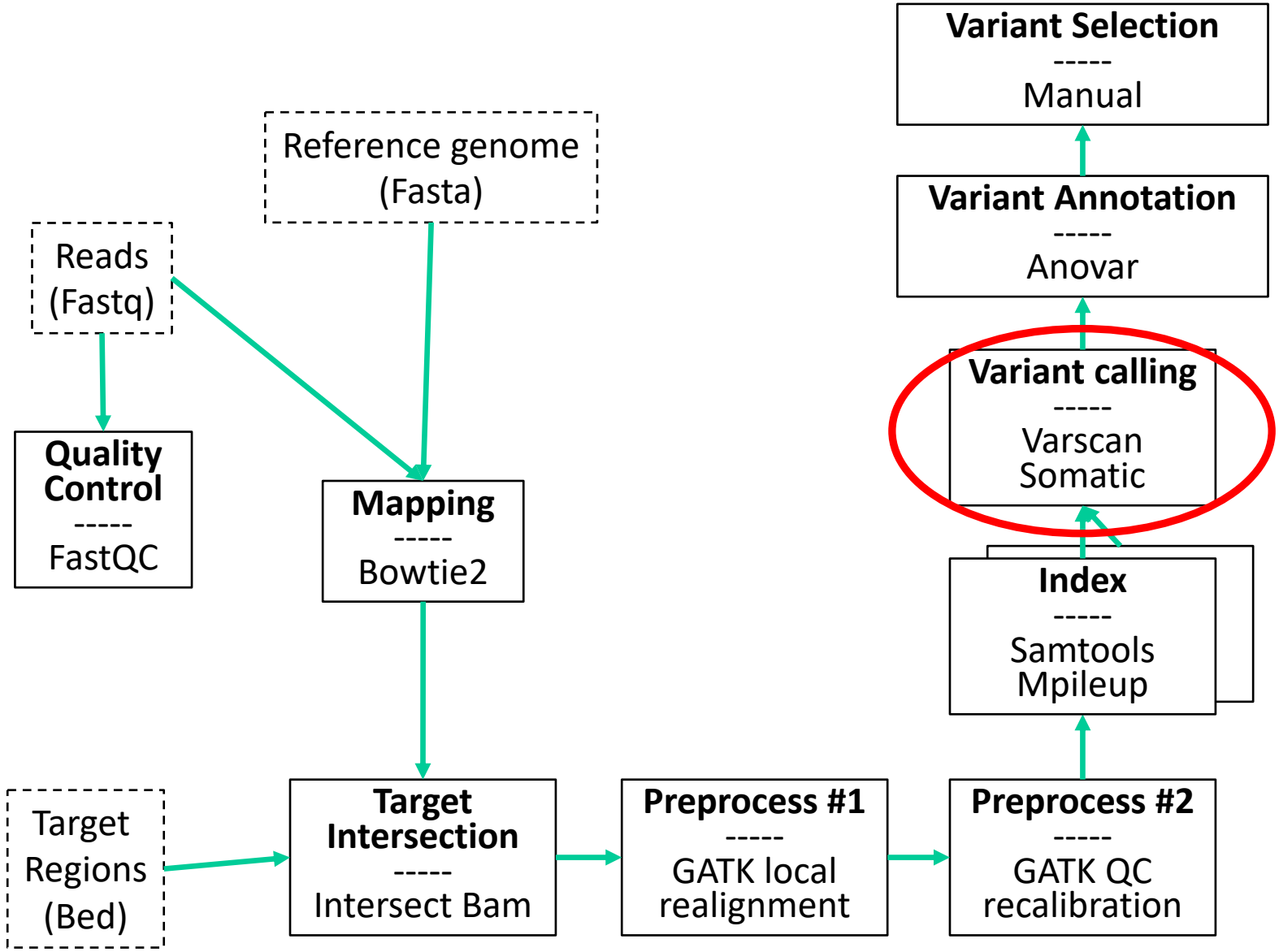
Analyse du PileUp



Producing the PileUp file

- samtools mpileup


Repeat for tumor & normal
Reference genome (Fasta)
Reads (Fastq)
Quality Control
FastQC
Mapping
BWA
Target Regions (Bed)
Target Intersection
Intersect Bam
Preprocess #1
GATK local realignment
Preprocess #2
GATK QC recalibration
Index
Samtools Mpileup
Variant calling
Varscan Somatic
Variant Annotation
Anovar
Variant Selection
Manual


Variant Selection

Manual
Variant Annotation

Anovar
Variant calling

Varscan
Somatic
Index

Samtools
Mpileup
Preprocess #2

GATK QC
recalibration
Preprocess #1

GATK local
realignment
Target
Intersection

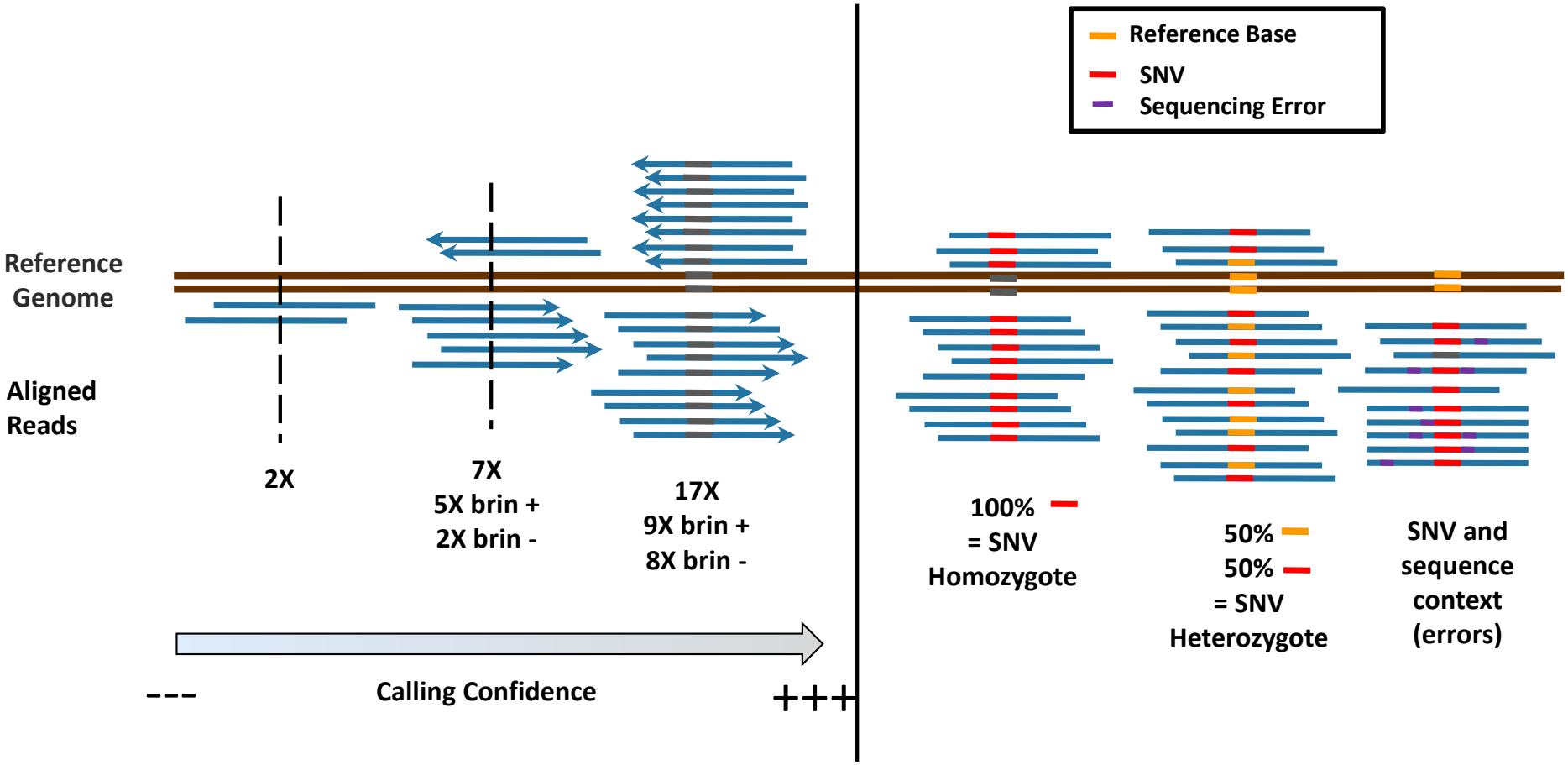
Intersect Bam
Target
Regions
(Bed)
Quality
Control

FastQC
Mapping

Bowtie2
Reads
(Fastq)
Reference genome
(Fasta)

Depth of Coverage

Depth of Coverage = number of reads supporting one position
ex: 1X, 5X, 100X... >1000X



Variant Calling

- Factors to consider when calling a SNV:
  - Base call qualities of each supporting base (base quality)
  - Proximity to small indels, or homopolymer run
  - Mapping qualities of the reads supporting the SNP
  - Sequencing **depth**: >=30x for constit ; >=100 for tumor
  - Position of SNV within reads: Higher error rate at the reads ends
  - Look at strand bias (SNVs supported by only one strand are more likely to be artifactual)
  - **Allelic frequency**: Tumor cellularity will reduce the % of an heterozygous variant
- Higher stringency when calling indels (and Sanger validation often needed)

VarScan2

- Mutation caller written in **Java** (no installation required) working with **Pileup files** of Targeted, Exome, and Whole-Genome sequencing data (DNAseq or RNAseq)

- **Multi-platforms**: Illumina, SOLiD, Life/PGM, Roche/454

- Detection of different kinds of Germline SNVs/Indels (classical mode):
  - Variants in individual samples
  - Multi-sample variants **shared or private** in multi-sample datasets

- VarScan specificity is to be able to work with **Tumor/Normal pairs (somatic mode):**
  - Somatic and germline mutation, LOH events in tumor-normal pairs
  - Somatic copy number alterations (CNAs) in tumor-normal exome data

VarScan2 Performance

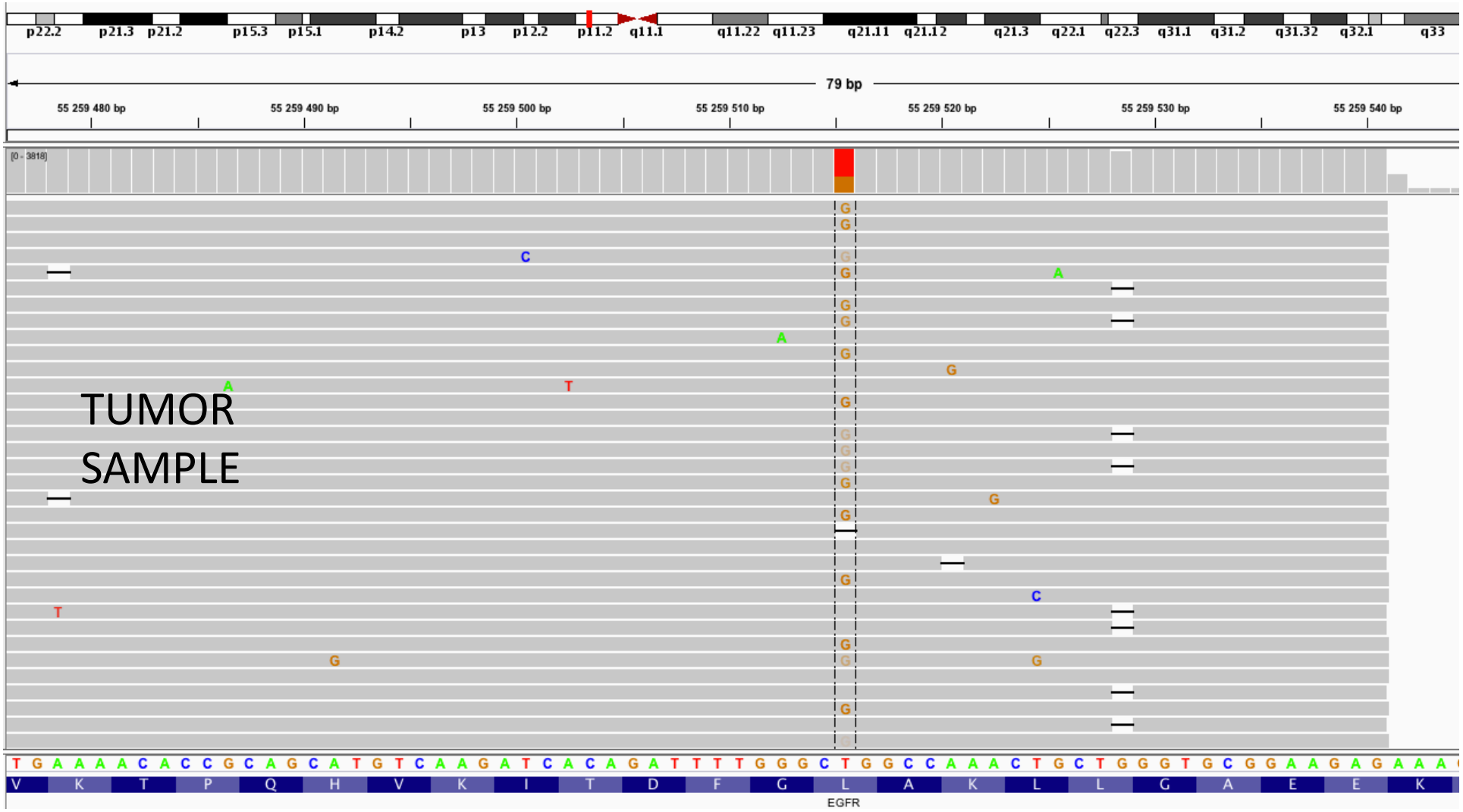
- VarScan uses a robust **heuristic/statistic** approach to call variants that meet desired thresholds for read depth, base quality, variant allele frequency, and statistical significance

- Stead *et* al. (2013) compared 3 different **somatic callers** : MuTect, Strelka, VarScan2
  - **VarScan2 performed best** overall with sequencing depths of 100x, 250x, 500x and 1000x required to accurately identify variants present at 10%, 5%, 2.5% and 1% respectively

- Other widely used tool: **GATK**

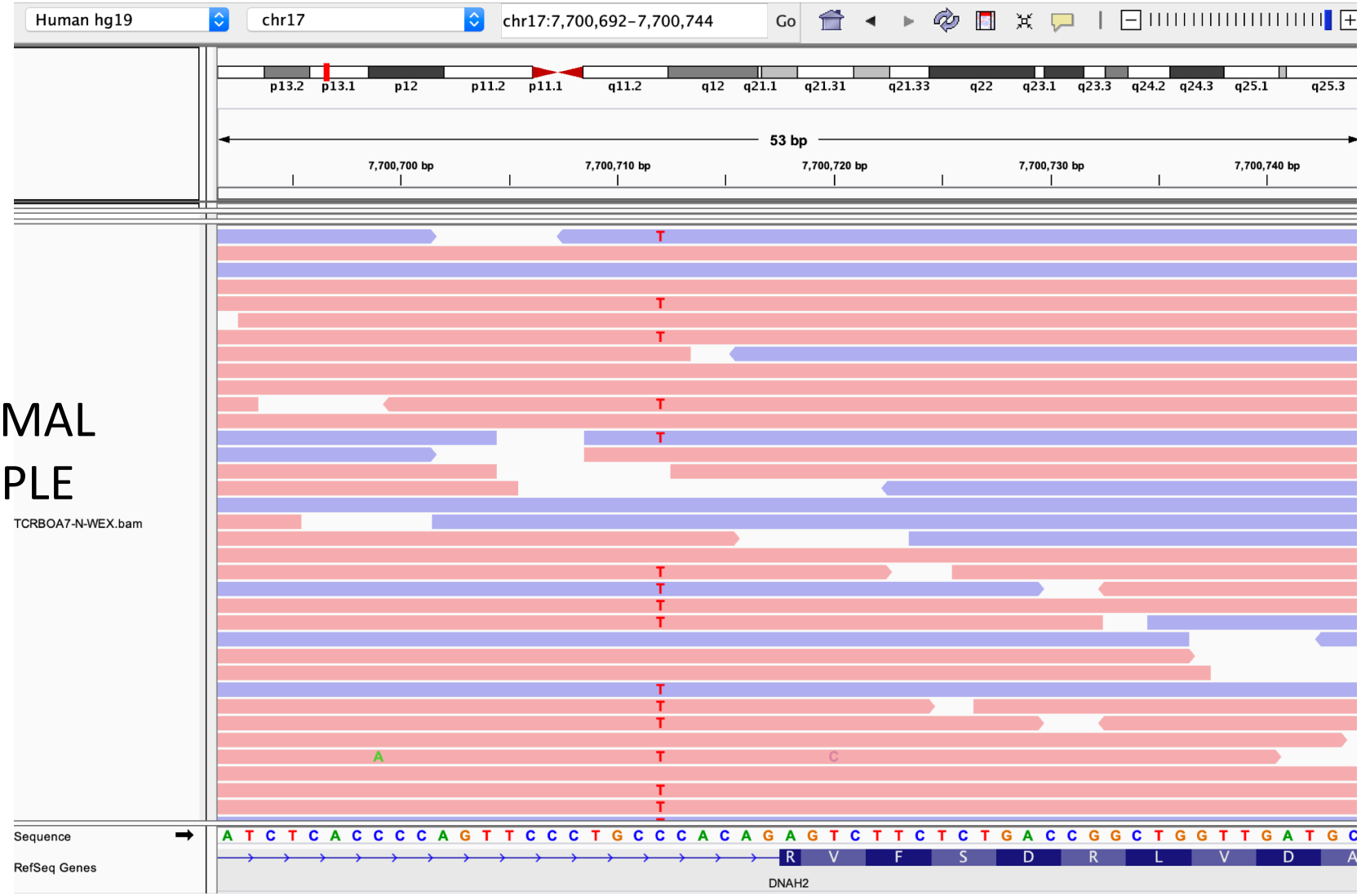
Criteria for somatic calling

Un variant tumoral: polymorphisme ou mutation?



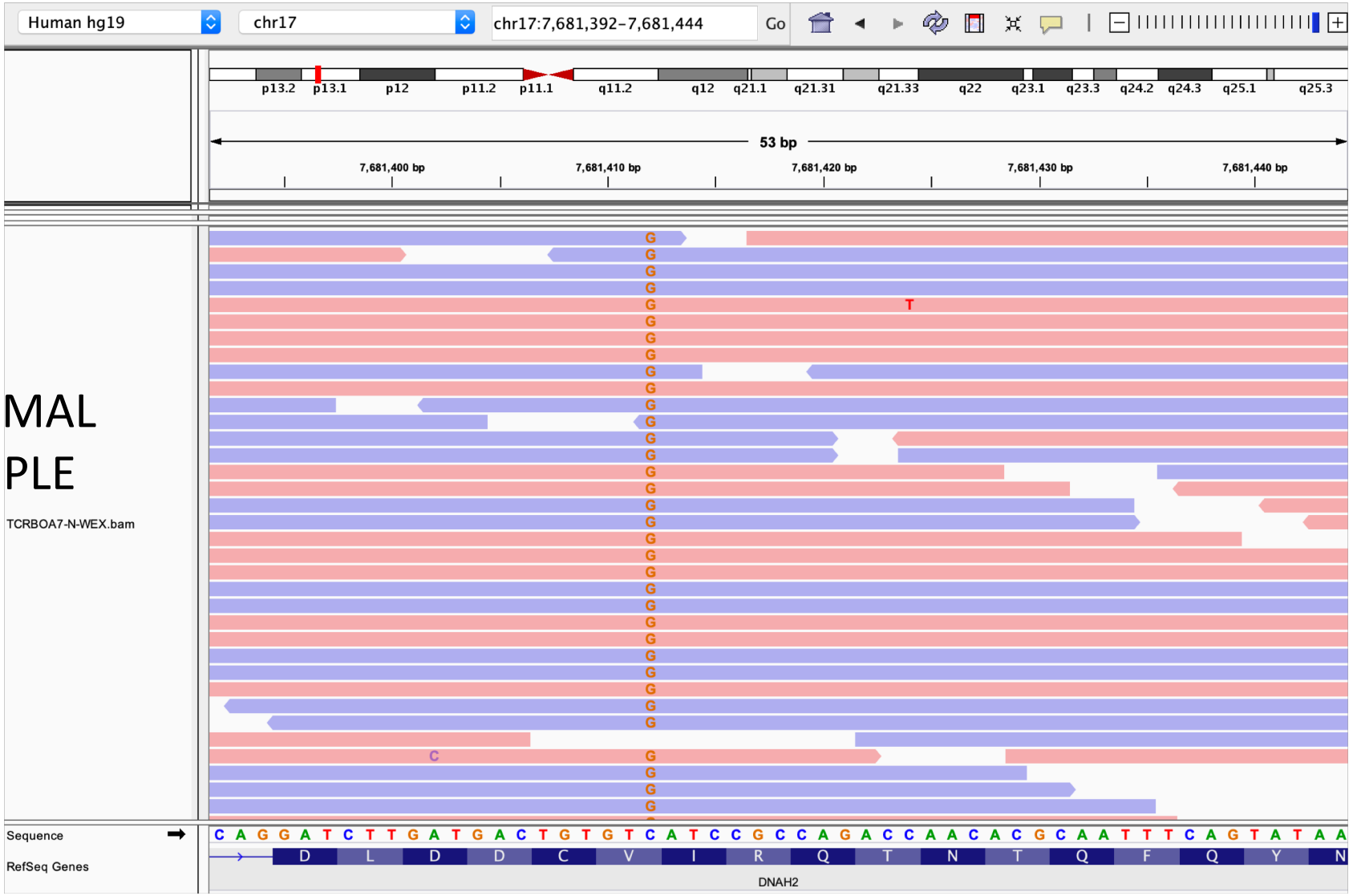
Un polymorphisme (SNP) hétérozygote

NORMAL SAMPLE

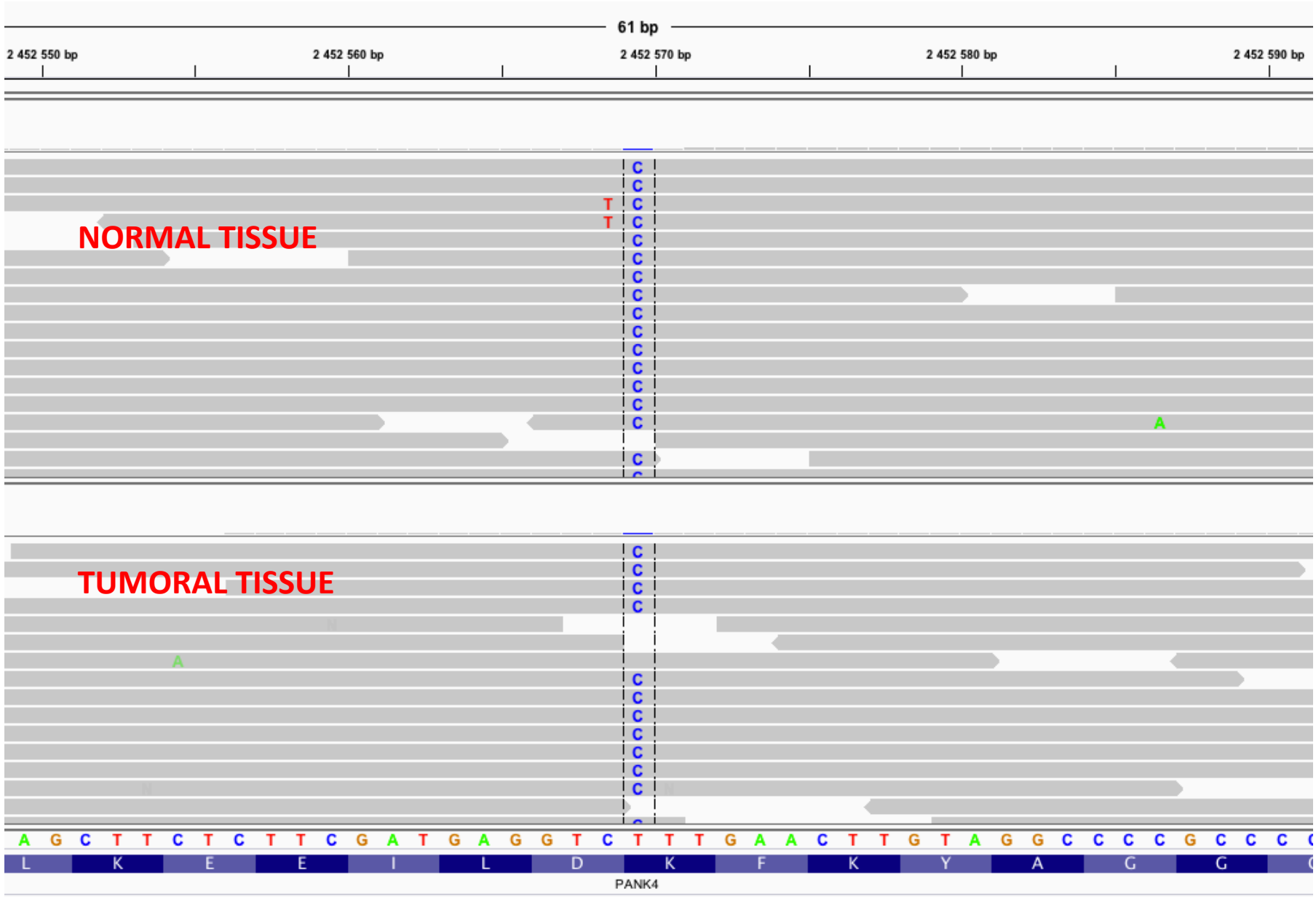


Un polymorphisme (SNP) homozygote

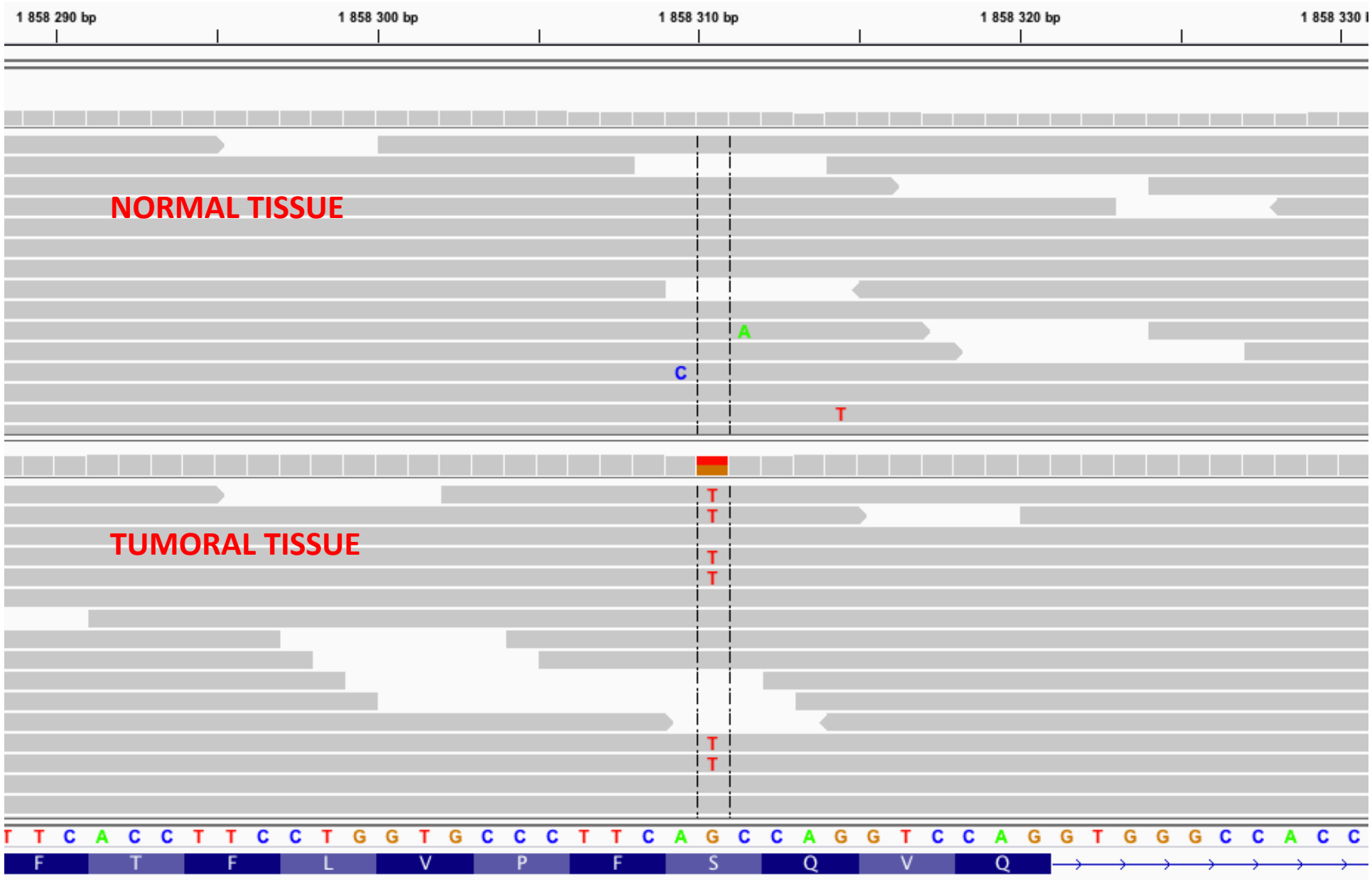
NORMAL SAMPLE



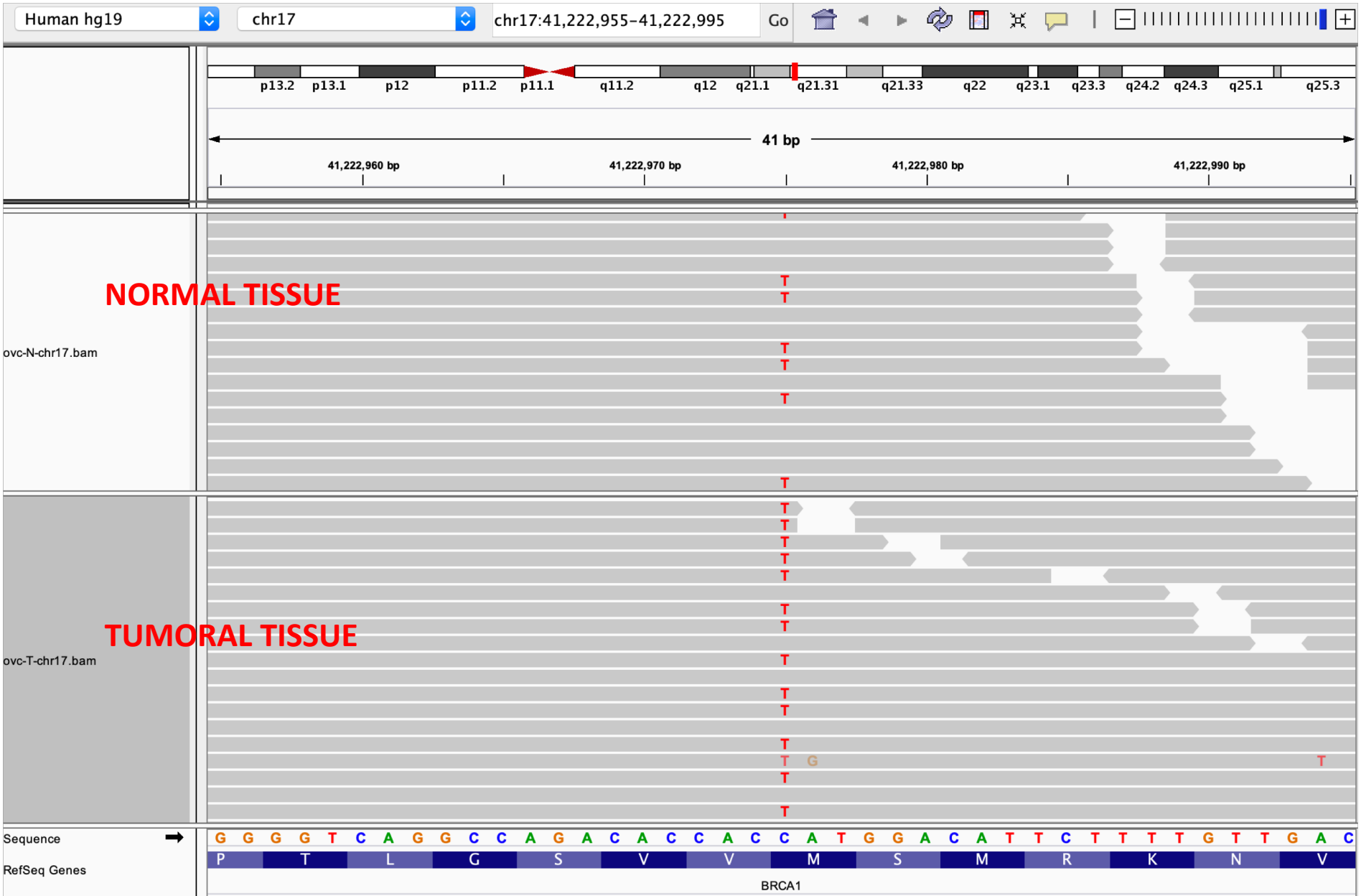
Un polymorphisme vu dans N et T



Une mutation somatique



Une LOH (loss of heterozygocity)



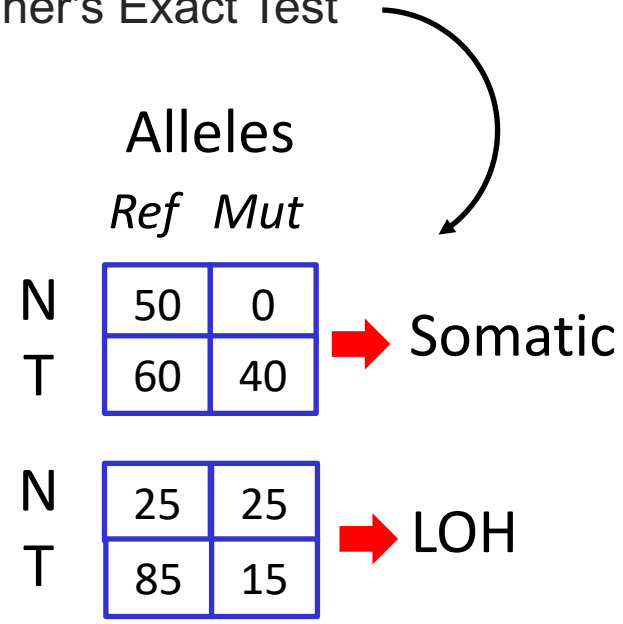
La fraction allélique

- Vocabulaire:
  - Germline/population: Allelic **frequency**, MAF (minor allele frequency). Par ex. dans données 1000Genomes.
  - Somatic: Allelic **fraction** (mais souvent on utilise VAF or BAF: variant allele frequency)
- Où trouver l'info?
  - Colonne info#AF dans VCF

Varscan's Somatic P-value

Variant Calling and Comparison
At every position where both normal and tumor have sufficient coverage, a comparison is made. First, normal and tumor are called independently using the germline consensus calling functionality. Then, their genotypes are compared by the following algorithm:

If **tumor does not match normal:**
Calculate significance of allele frequency difference by Fisher's Exact Test
If difference is significant (p-value < threshold):
If normal matches reference
==> Call Somatic
Else If normal is heterozygous
==> Call LOH
Else normal and tumor are variant, but different
==> Call IndelFilter or Unknown
If difference is not significant:
==> Call Germline



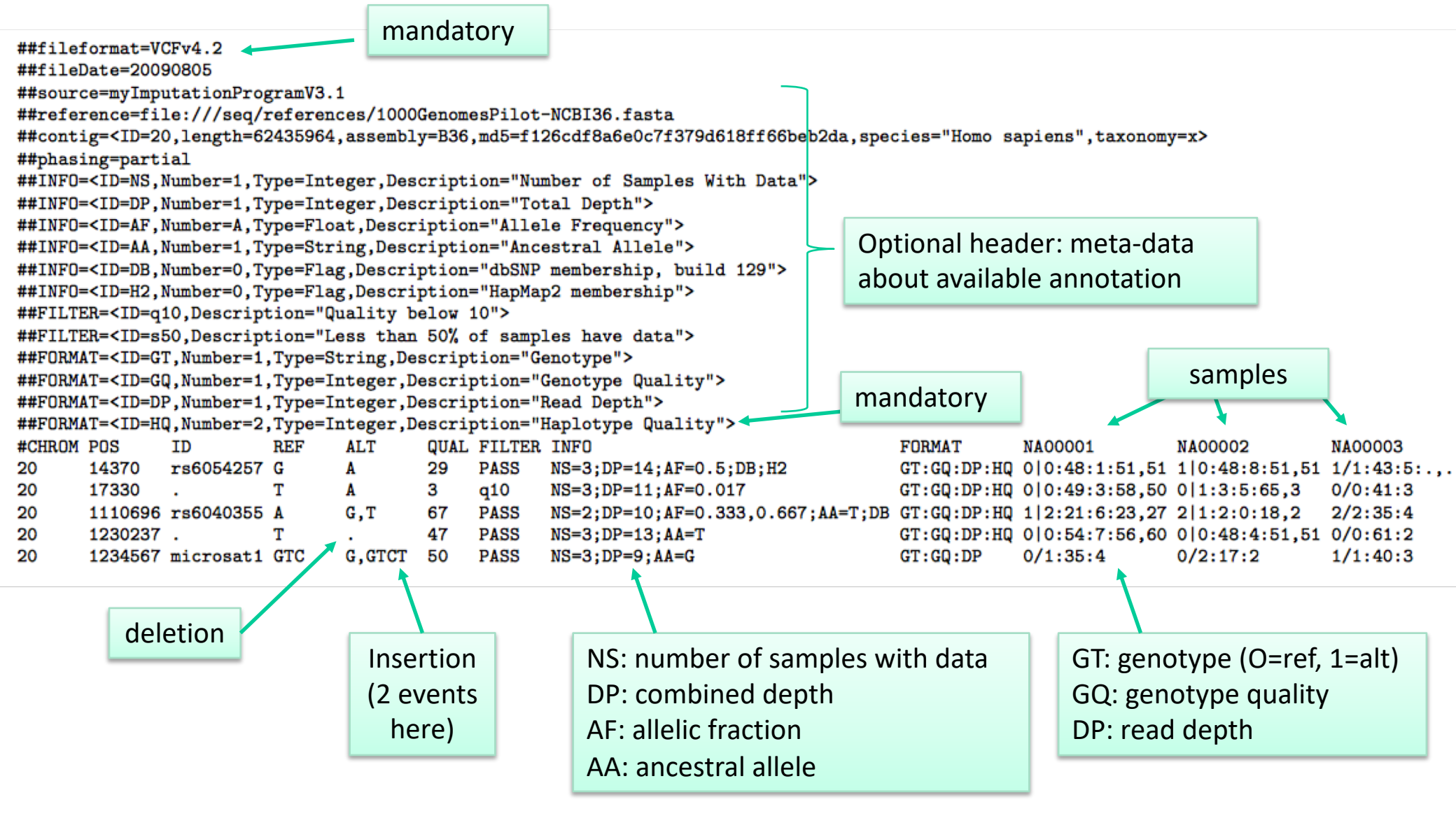
VarScan somatic

Parameters:

- Min-var-freq: minimal allelic frequency to call a variant (10% here)
- Min-coverage: minimum coverage to call a variant (in normal and tumor and combined)
- Tumor and normal purity: cellularity of your sample
- P-value threshold to call a heterozygote
- P-value threshold to call a somatic site

➢ 2 output files: SNVs & Indels in VCF format

Format VCF


mandatory
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
Optional header: meta-data about available annotation
mandatory
samples
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
deletion
Insertion (2 events here)
NS: number of samples with data
DP: combined depth
AF: allelic fraction
AA: ancestral allele
GT: genotype (O=ref, 1=alt)
GQ: genotype quality
DP: read depth

VarScan Tabulated Format

Chrom	Position	Ref	Cons	Reads1	Reads2	VarFreq	Strands 1	Strands 2	Qual1	Qual2	Pvalue	Map Qual1	Map Qual2	R1 +	R1 -	R2 +	Rs2 -	Alt
chr12	113348849	C	Y	31	30	**49.18%**	2	2	27	27	0.98	1	1	19	12	25	5	**T**
chr12	113354329	G	R	72	2	**2.70%**	2	2	31	26	0.98	1	1	48	24	1	1	**A**
chr12	113357193	G	A	2	72	**97.30%**	1	2	28	24	0.98	1	1	2	0	45	27	**A**
chr12	113357209	G	A	0	77	**100%**	0	2	0	29	0.98	0	1	0	0	51	26	**A**

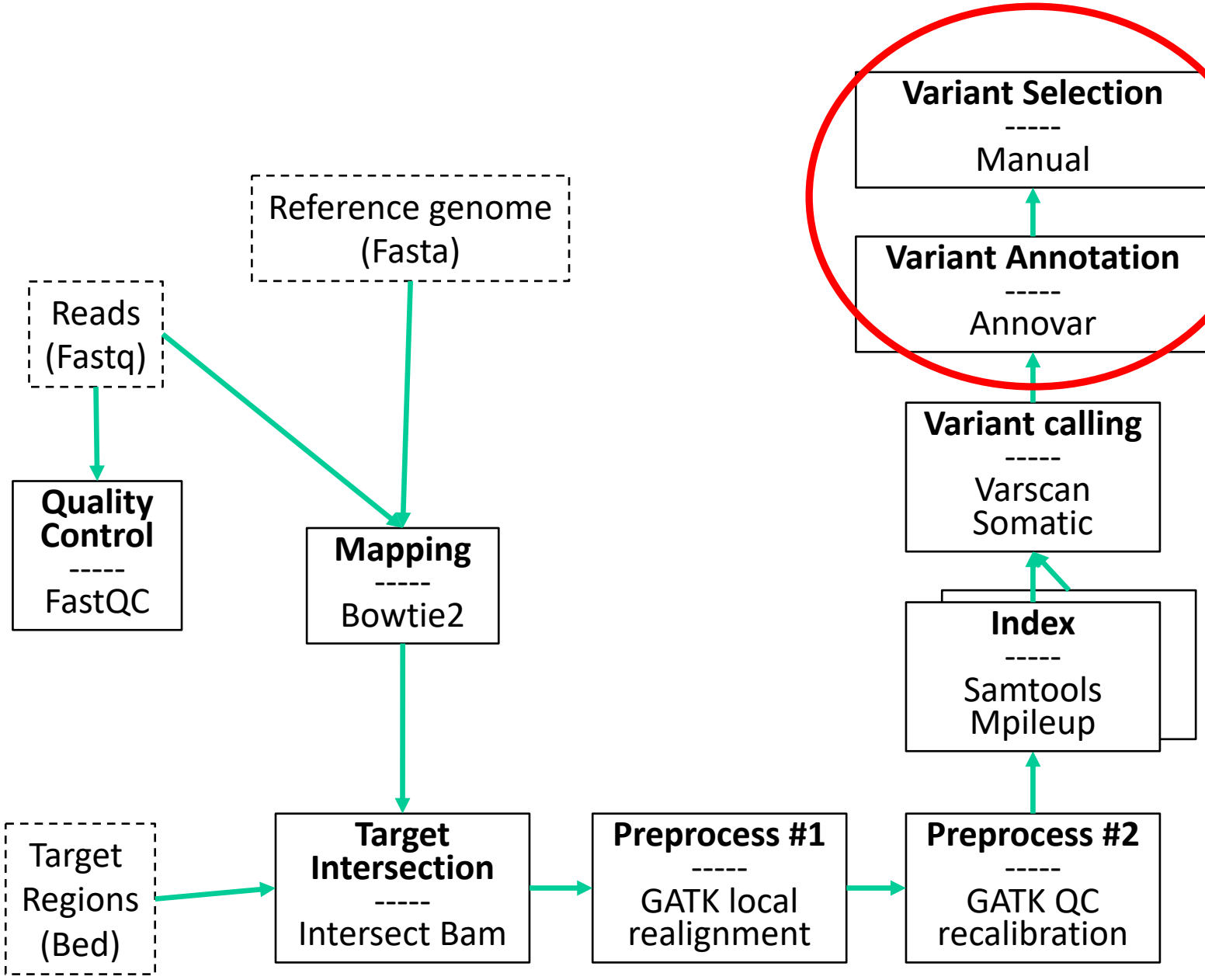
Cons : Consensus Genotype of Variant Called (IUPAC code):

M -> A or C	Y -> C or T	D -> A or G or T	W -> A or T	V -> A or C or G
R -> A or G	K -> G or T	B -> C or G or T	S -> C or G	H -> A or C or T

MAF format

Mutation Annotation Format (MAF) is a tab-delimited text file with aggregated mutation information from VCF Files and are generated on a project-level.

Column	Description
1 - Hugo_Symbol	HUGO symbol for the gene (HUGO symbols are always in all caps). "Unknown" is used for regions that do not correspond to a gene
2 - Entrez_Gene_Id	Entrez gene ID (an integer). "0" is used for regions that do not correspond to a gene region or Ensembl ID
3 - Center	One or more genome sequencing center reporting the variant
4 - NCBI_Build	The reference genome used for the alignment (GRCh38)
5 - Chromosome	The affected chromosome (chr1)
6 - Start_Position	Lowest numeric position of the reported variant on the genomic reference sequence. Mutation start coordinate
7 - End_Position	Highest numeric genomic position of the reported variant on the genomic reference sequence. Mutation end coordinate
8 - Strand	Genomic strand of the reported allele. Currently, all variants will report the positive strand: '+'
9 - Variant_Classification	Translational effect of variant allele
10 - Variant_Type	Type of mutation. TNP (tri-nucleotide polymorphism) is analogous to DNP (di-nucleotide polymorphism) but for three consecutive nucleotides. ONP (oligo-nucleotide polymorphism) is analogous to TNP but for consecutive runs of four or more (SNP, DNP, TNP, ONP, INS, DEL, or Consolidated)
11 - Reference_Allele	The plus strand reference allele at this position. Includes the deleted sequence for a deletion or "-" for an insertion
12 - Tumor_Seq_Allele1	Primary data genotype for tumor sequencing (discovery) allele 1. A "-" symbol for a deletion represents a variant. A "-" symbol for an insertion represents wild-type allele. Novel inserted sequence for insertion does not include flanking reference bases
13 - Tumor_Seq_Allele2	Tumor sequencing (discovery) allele 2
14 - dbSNP_RS	The rs-IDs from the dbSNP database, "novel" if not found in any database used, or null if there is no dbSNP record, but it is found in other databases
15 - dbSNP_Val_Status	The dbSNP validation status is reported as a semicolon-separated list of statuses. The union of all rs-IDs is taken when there are multiple
16 - Tumor_Sample_Barcode	Aliquot barcode for the tumor sample
17 - Matched_Norm_Sample_Barcode	Aliquot barcode for the matched normal sample
18 - Match_Norm_Seq_Allele1	Primary data genotype. Matched normal sequencing allele 1. A "-" symbol for a deletion represents a variant. A "-" symbol for an insertion represents wild-type allele. Novel inserted sequence for insertion does not include flanking reference bases (cleared in somatic MAF)
19 - Match_Norm_Seq_Allele2	Matched normal sequencing allele 2


Variant Selection

Manual
Variant Annotation

Annovar
Reference genome
(Fasta)
Reads
(Fastq)
Variant calling

Varscan
Somatic
Quality
Control

FastQC
Mapping

Bowtie2
Index

Samtools
Mpileup
Target
Regions
(Bed)
Target
Intersection

Intersect Bam
Preprocess #1

GATK local
realignment
Preprocess #2

GATK QC
recalibration

Different types of SNVs

- SNVs and short indels qualification:
  Intergenic
  Intronic
  Cis-regulatory
  Splice sites
  Frameshift or not
  Synonymous or not
  Benign or damaging etc...

- « Interesting » SNVs:
  Non-synonymous + highly deleterious + somatically acquired

Resources dedicated to human genetic variation

- dbSNP and 1000-genomes
  - Population-scale DNA polymorphisms
- COSMIC
  - Catalogue Of Somatic Mutations In Cancer
- Non synonymous SNVs predictions
  - SIFT, Polyphen2 (damaging impact)... PhyloP, GERP++ (conservation)

Annovar

Annovar annotates SNVs and Indels
Input: Multi-sample VCF (Tumor + Normal samples)
Information generated:

- RefGene: Gene & Function & Aminoacid Change (HGVS format: c.A155G ; p.Lys45Arg)
- 1000g2012apr_all: Minor Allele Frequency for all ethnies
- ESP6500: Exome Sequencing Project
- Ljb_all : predictions (**SIFT**, **Polyphen2**, LRT, MutationTaster, PhyloP, GERP++)

✧ Tabulated output file:

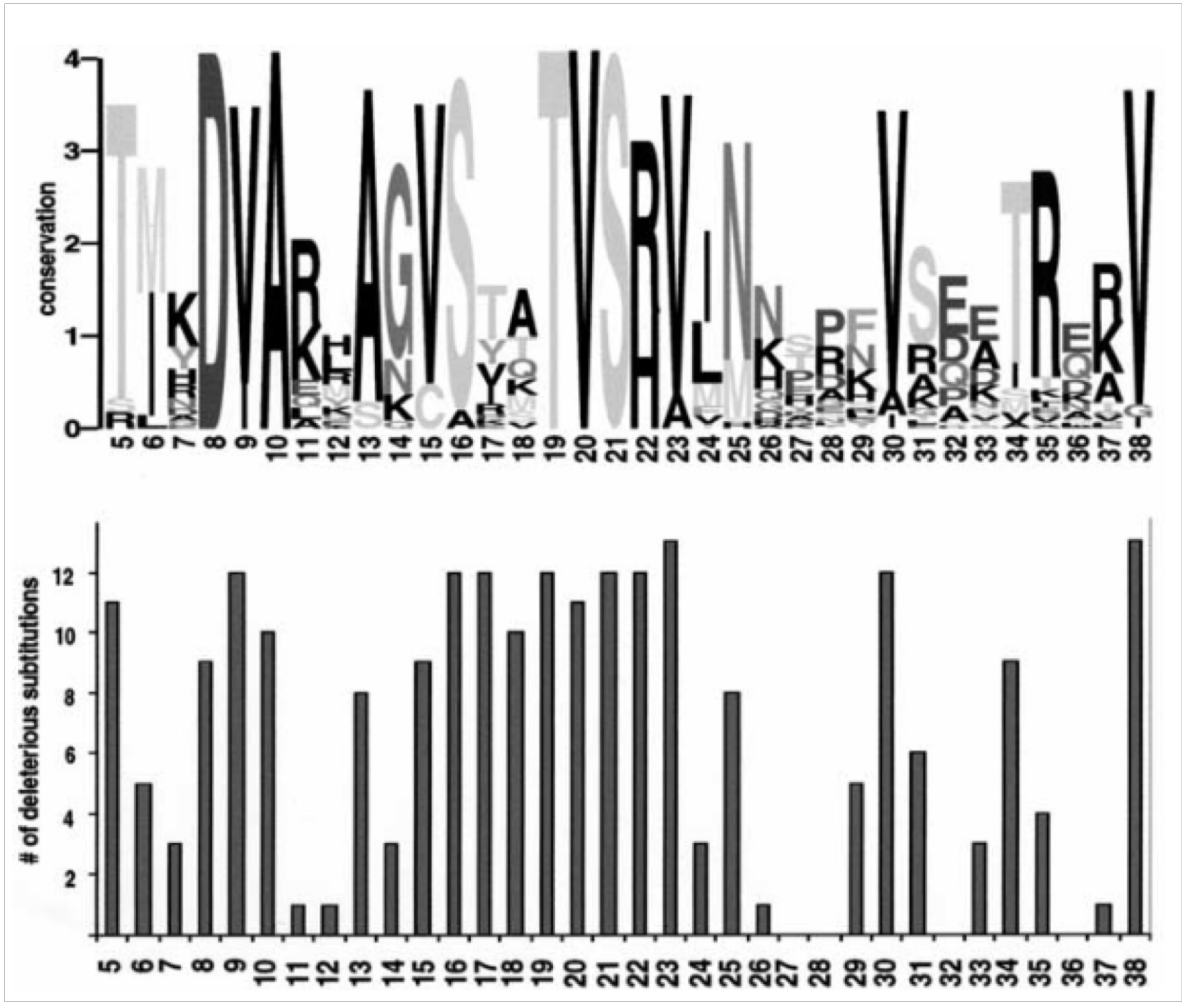
Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	ExonicFunc.refGene	AAChange.refGene	1000g2012apr_all	snp137	cosmic68	esp6500_all	LJB_PhyloP	LJB_PhyloP_Pre
chr1	160251792	160251792	A	G	intronic	PEX19			NA				NA	
chr1	167082869	167082869	G	A	intronic	DUSP27			NA				NA	
chr1	167095163	167095163	G	C	exonic	DUSP27	nonsynonymous SNV	DUSP27:NM_001080426:exon5:c.G795C:p.E265D						
chr1	167095881	167095881	G	A	exonic	DUSP27	nonsynonymous SNV	DUSP27:NM_001080426:exon5:c.G1513A:p.A505T						
chr1	167097739	167097739	C	A	exonic	DUSP27	nonsynonymous SNV	DUSP27:NM_001080426:exon5:c.C3371A:p.T1124N						
chr1	214803969	214803969	G	C	exonic	CENPF	nonsynonymous SNV	CENPF:NM_016343:exon9:c.G1287C:p.K429N						

Sorting Intolerant From Tolerant

Ng & Henikoff, Genome Res. 2001

Utilise la conservation des domaines protéiques comme indication du caractère délétère d'une substitution

Classe en **D**(eleterious), **T**(olerant), **.**(unknown)



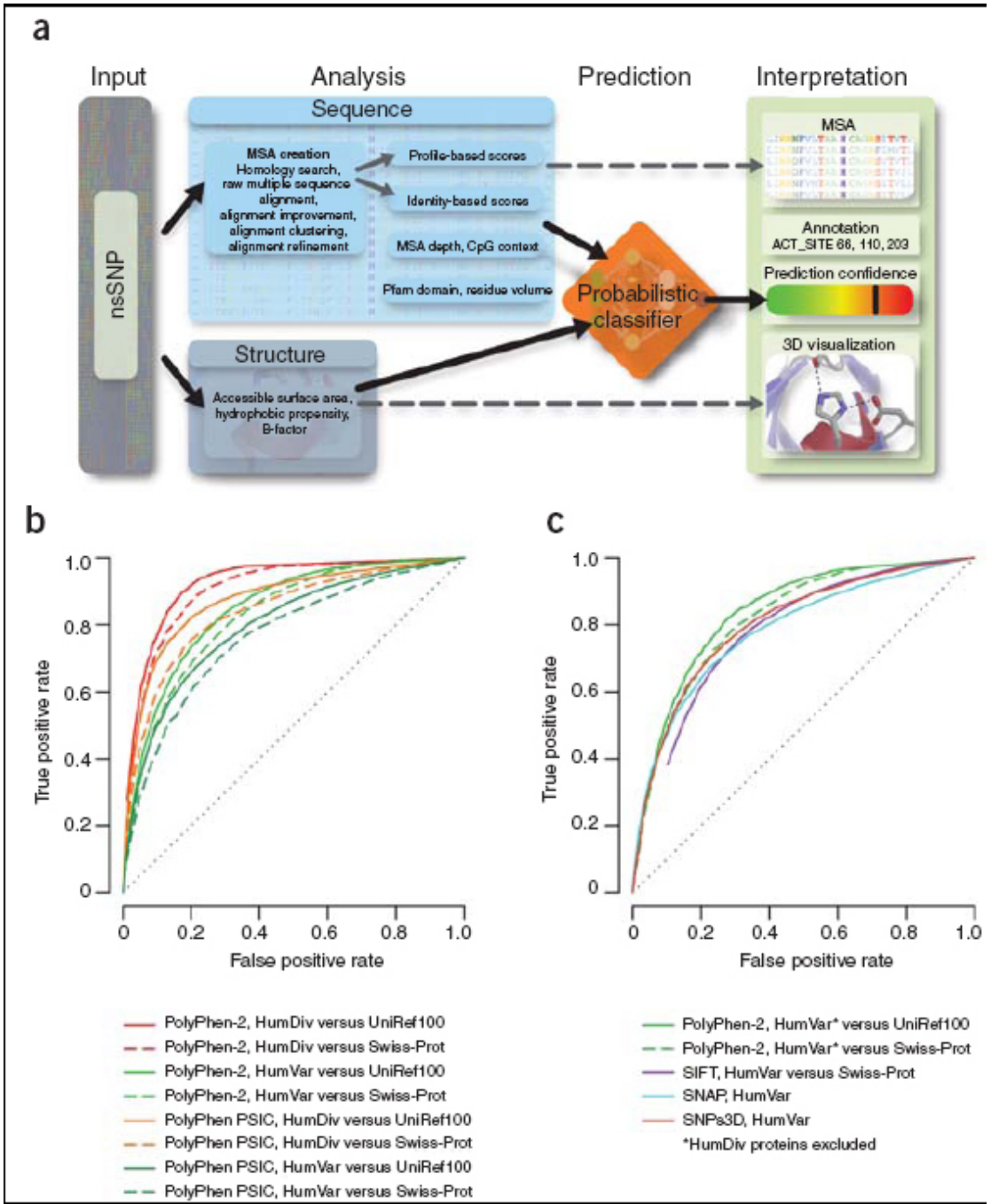
effet réél: LacI

PolyPhen2

Adzhubei et al. Nature Methods 2010

Probabilistic classifier: Estimates the probability of the missense mutation being damaging based on a combination of seq+struct properties.

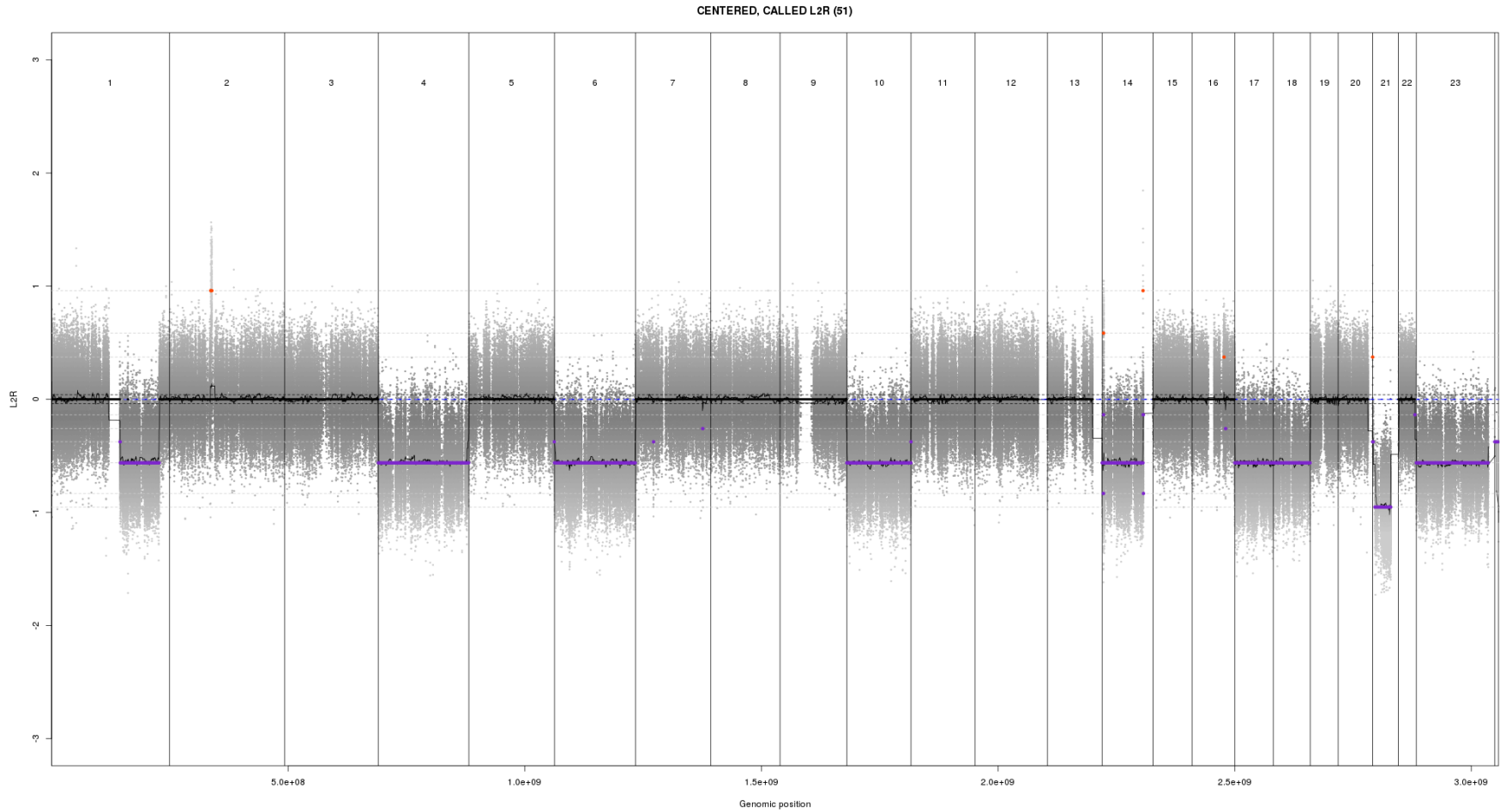
Classe en: **B**enign, **P**ossibly damaging, or probably **D**amaging



Coverage & Allelic Frequencies For CNV detection

Detection of copy-number variations

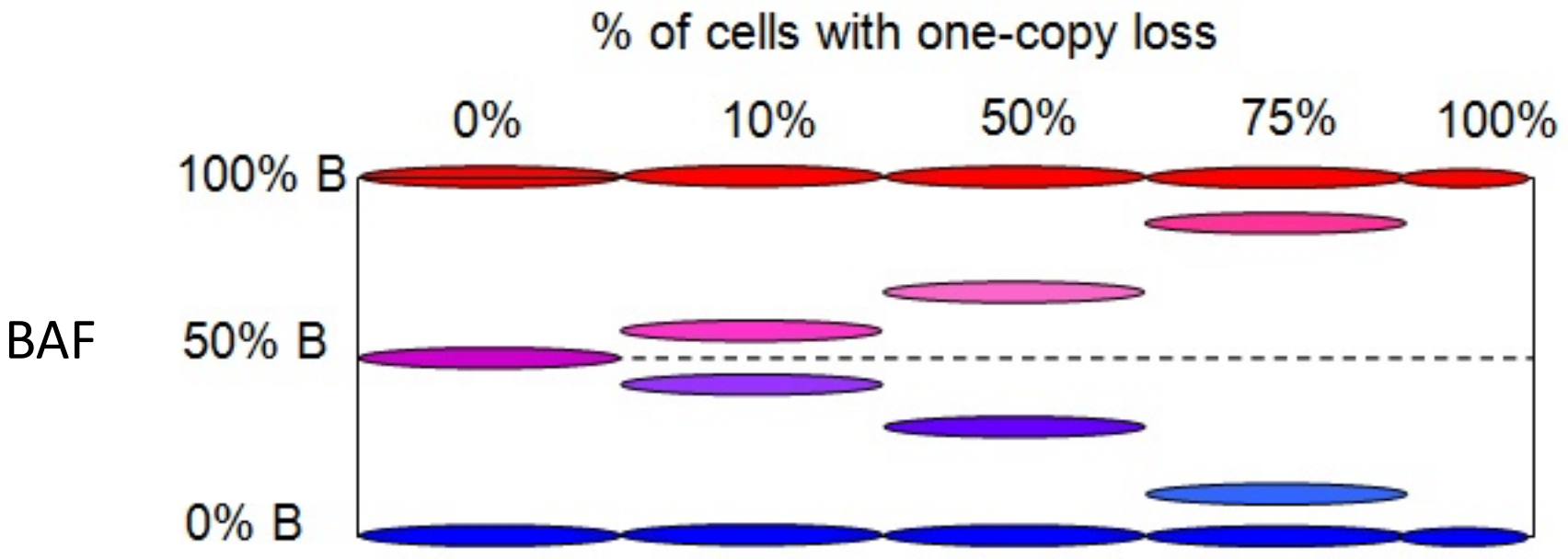
Are there any copy-number alteration (gain or loss of chomosomal regions, amplifications …) that could explain tumorigenesis ?



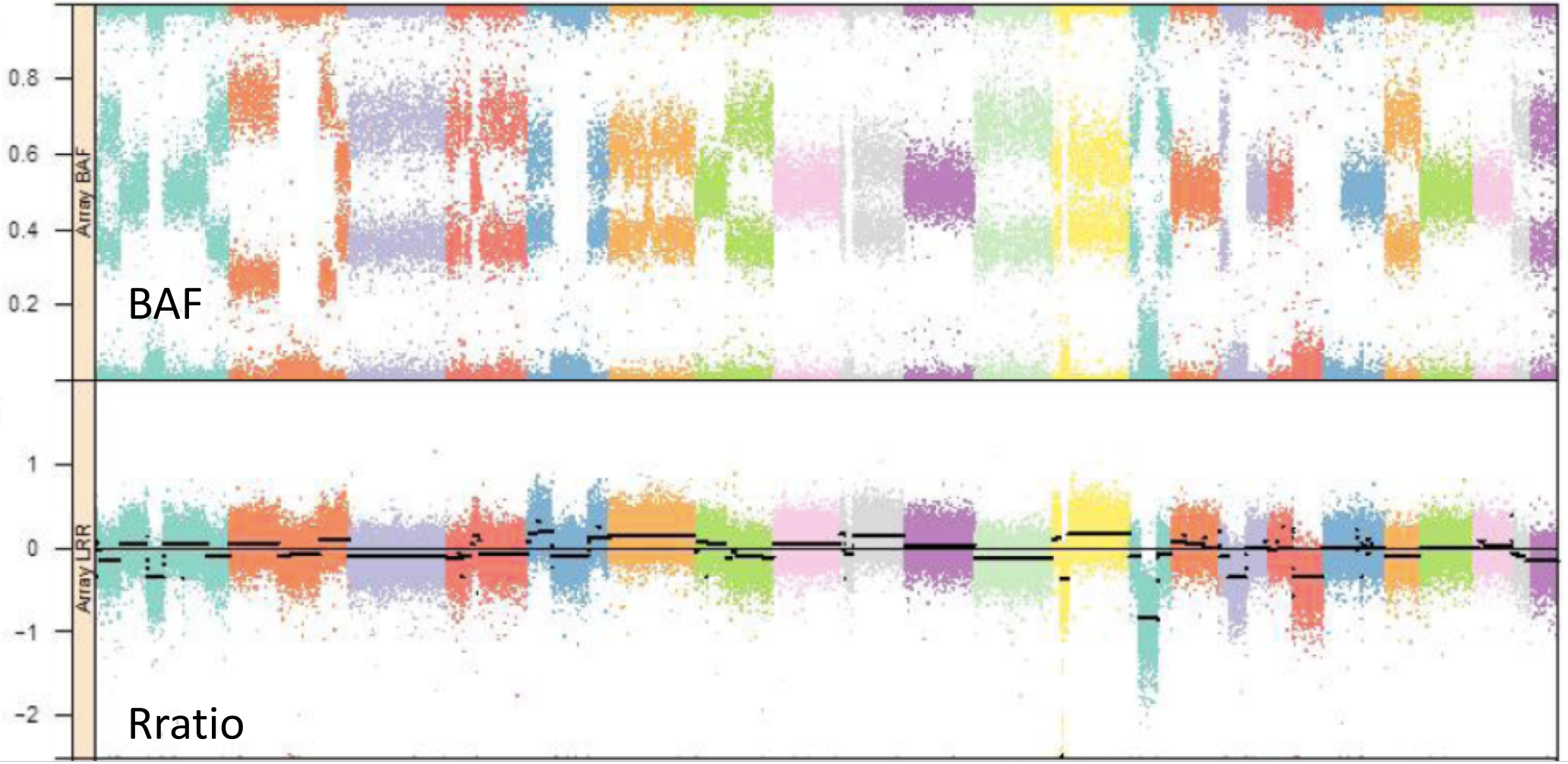
Fréquence ou fraction allélique

- Vocabulary:
  - In a population: Allelic **frequency**, MAF (minor allele frequency). For instance in 1000Genomes.
  - In tumor: Allelic **fraction** (often VAF or BAF are used: variant allele frequency)
- Where is it found?
  - Colonne info#AF dans VCF

Cellularité et Fraction Allélique



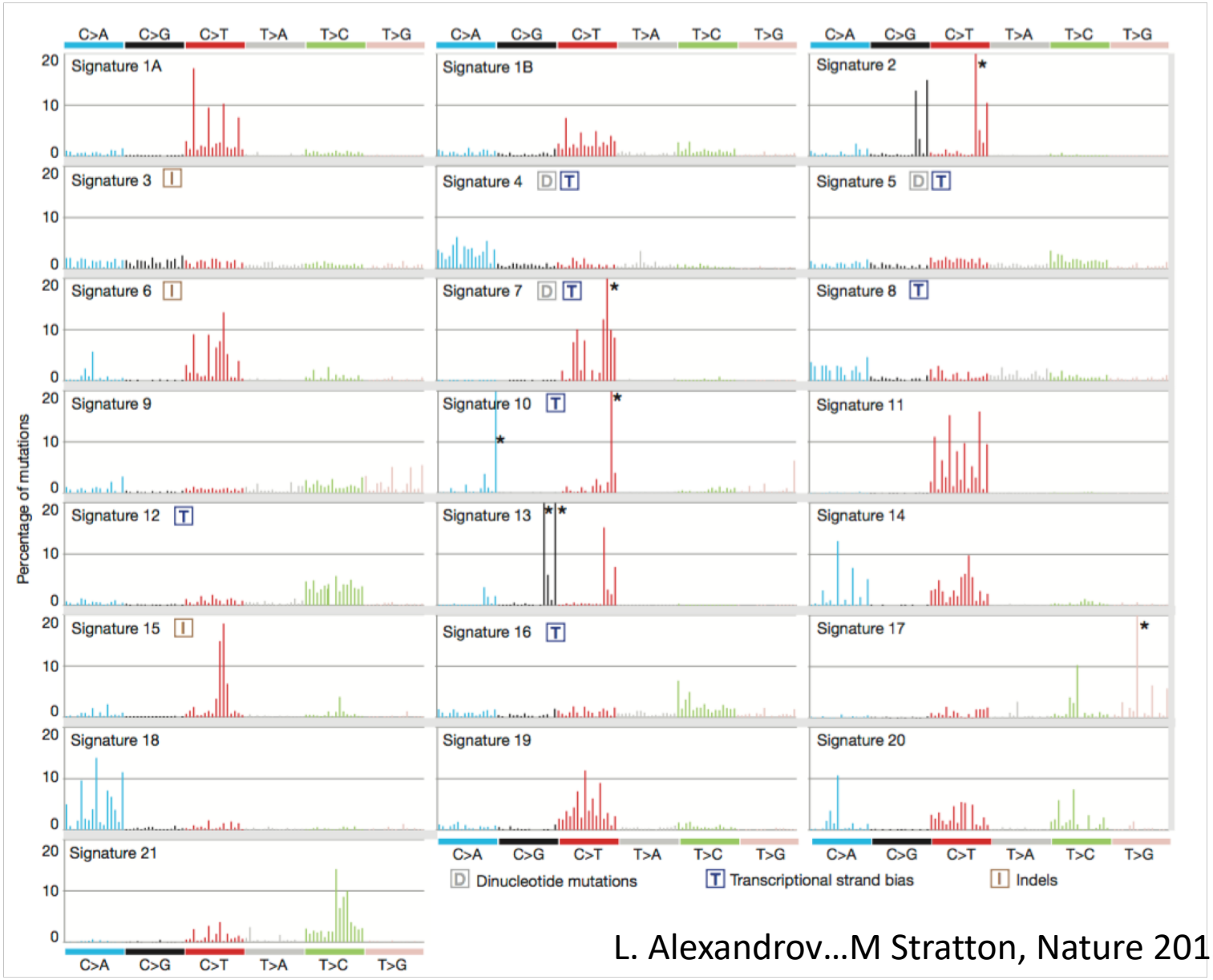
Segmentation et fréquence allélique



R ratio=utilisé en CGH, =couverture en NGS

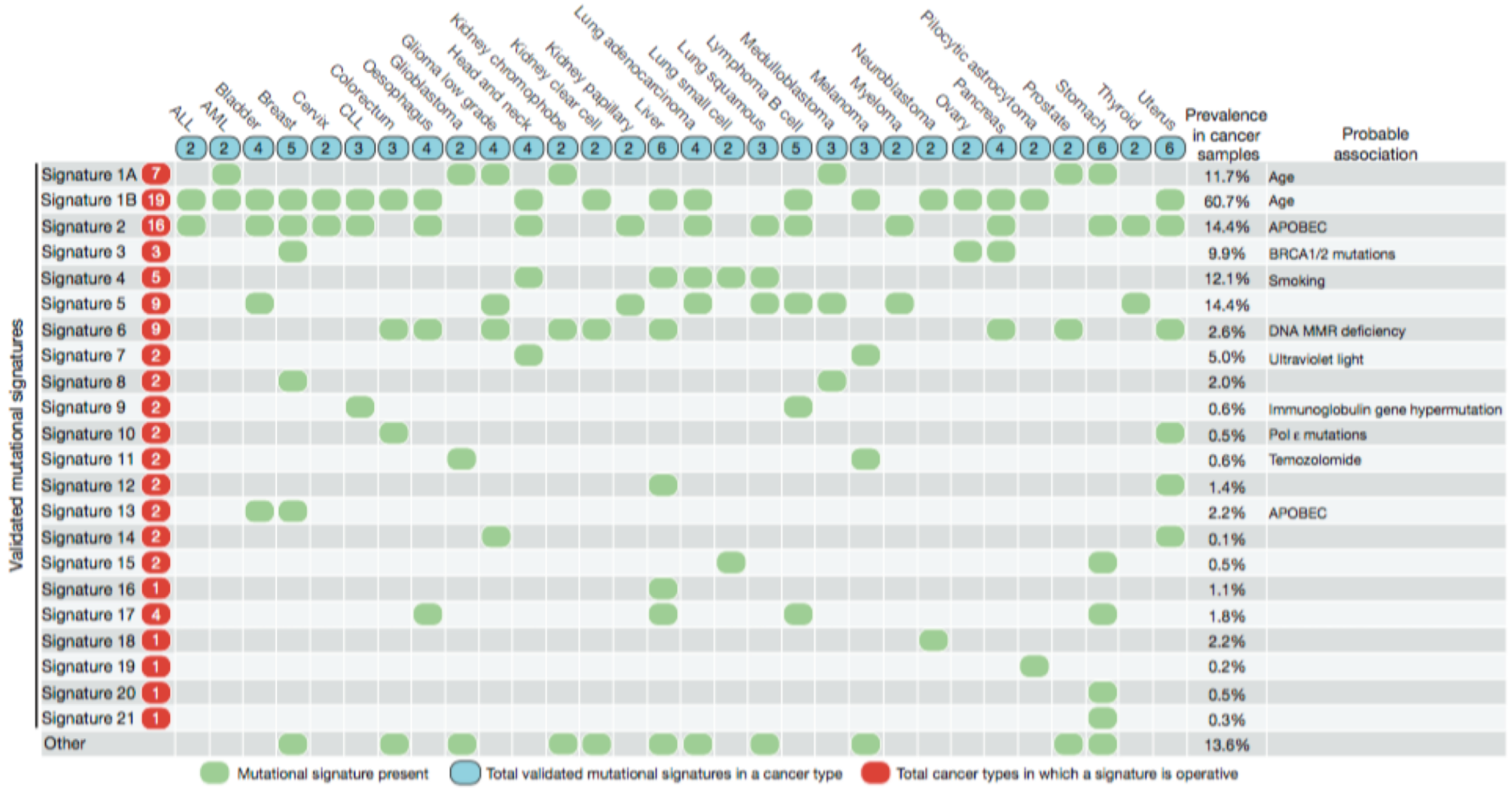
Scott et al. Gene 2014

Les Signatures Mutationnelles



L. Alexandrov…M Stratton, Nature 2013.

Signatures et origine des tumeurs



L. Alexandrov…M Stratton, Nature 2013.