

Recherche de variants génomiques en oncologie clinique

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

Technologies de recherche de variants

- Sanger
 - Toujours utilisé en consultation de génétique
- SNP arrays
 - ~1M SNP (recherche GWAS ou « 23&Me »)
- Panel de gènes
 - Une série d'exons d'intérêt (gènes de cancer= 100kb)
- WES (Exome)
 - Tous les exons du génome (30 Mb)
- WGS
 - Whole genome (3 Gb)

NGS

Variants par NGS: 3 approches

- Panel de gènes
 - Une série d'exons d'intérêt (gènes de cancer= 100kb)
- Exome
 - Tous les exons du génome (30-50 Mb)
- Whole genome
 - Le génome complet (3 Gb)

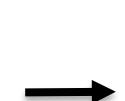
En génétique constitutionnelle

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...

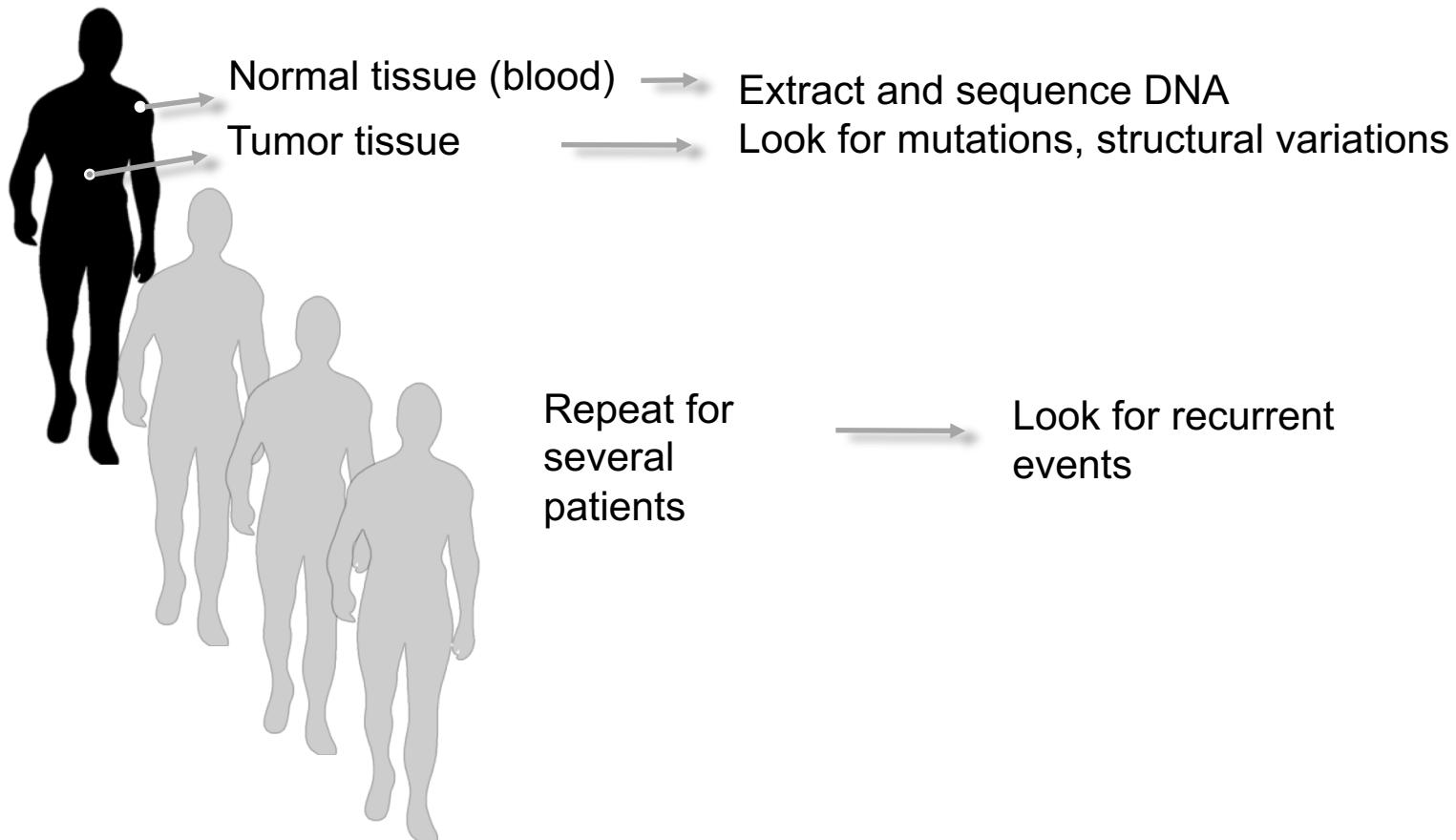
NGS dans le diagnostic de génétique familiale

Panel de gènes pour cibler:

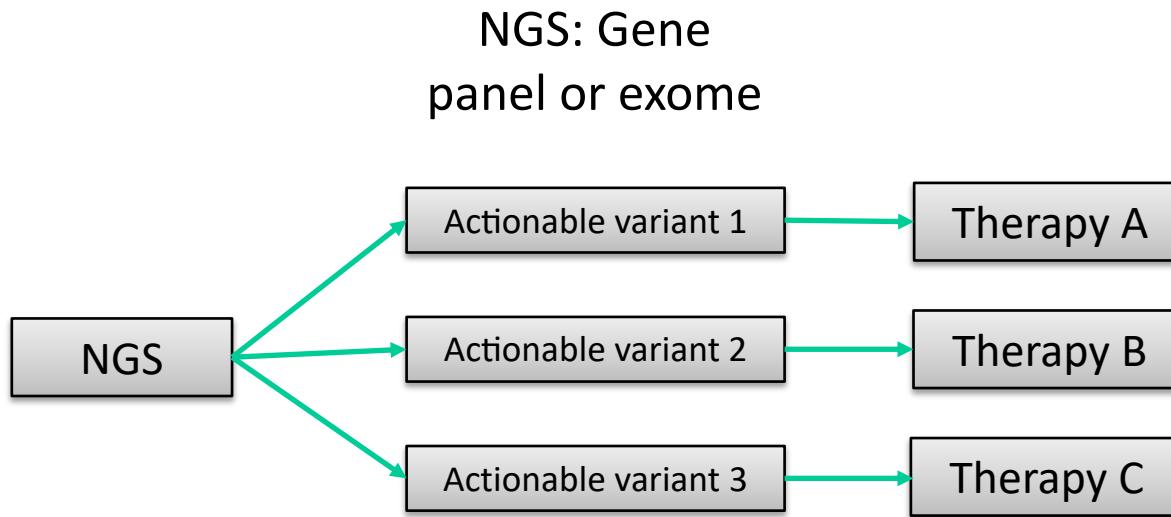
- BRCA1/2 (breast/ovary cancer)
- XPC, XPV.. (melanoma)
- ERCC1 (colorectal cancer)

En génétique somatique

Finding somatic mutations in the tumor genome

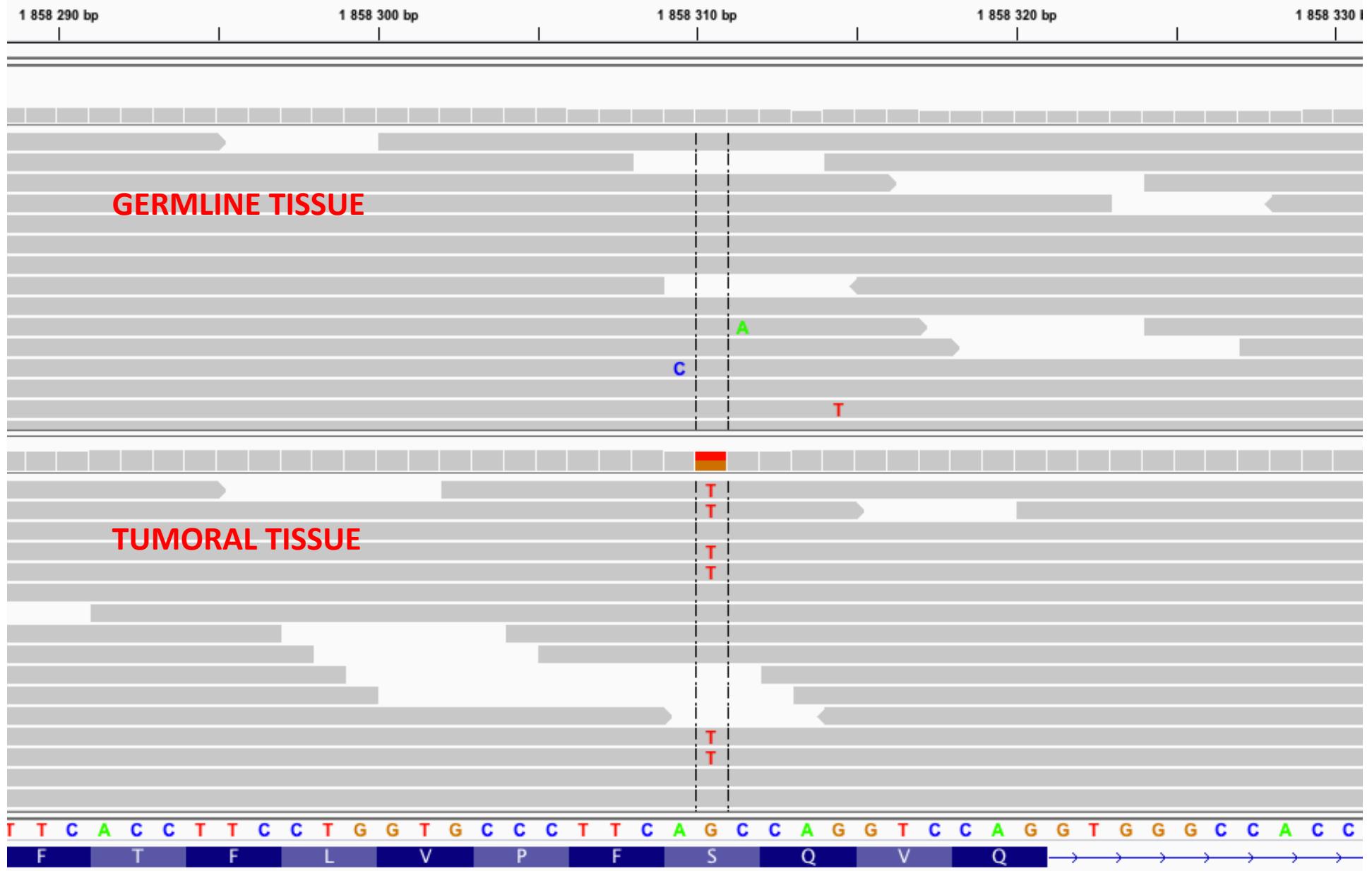


NGS for precision medicine

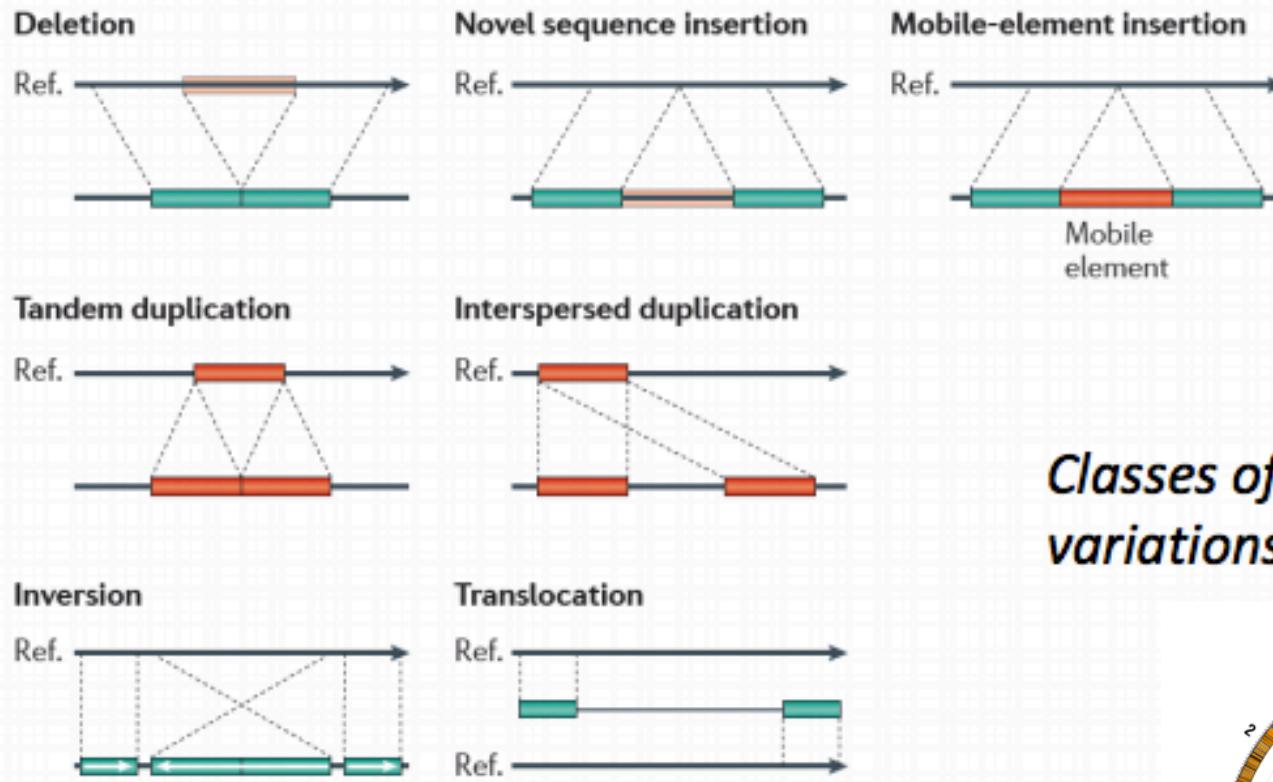


- Clinical trials: MOSCATO (GR), SAFIR (GR), SHIVA (Curie), ...
- Ipilimumab (anti-CTLA4), Nivolumab (anti-PD1), Trastuzumab (anti-HER2), Cetuximab (anti-EGFR)

Les mutations somatiques

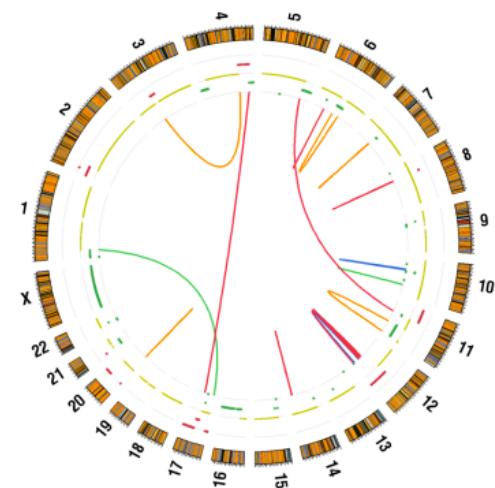


Variants structuraux



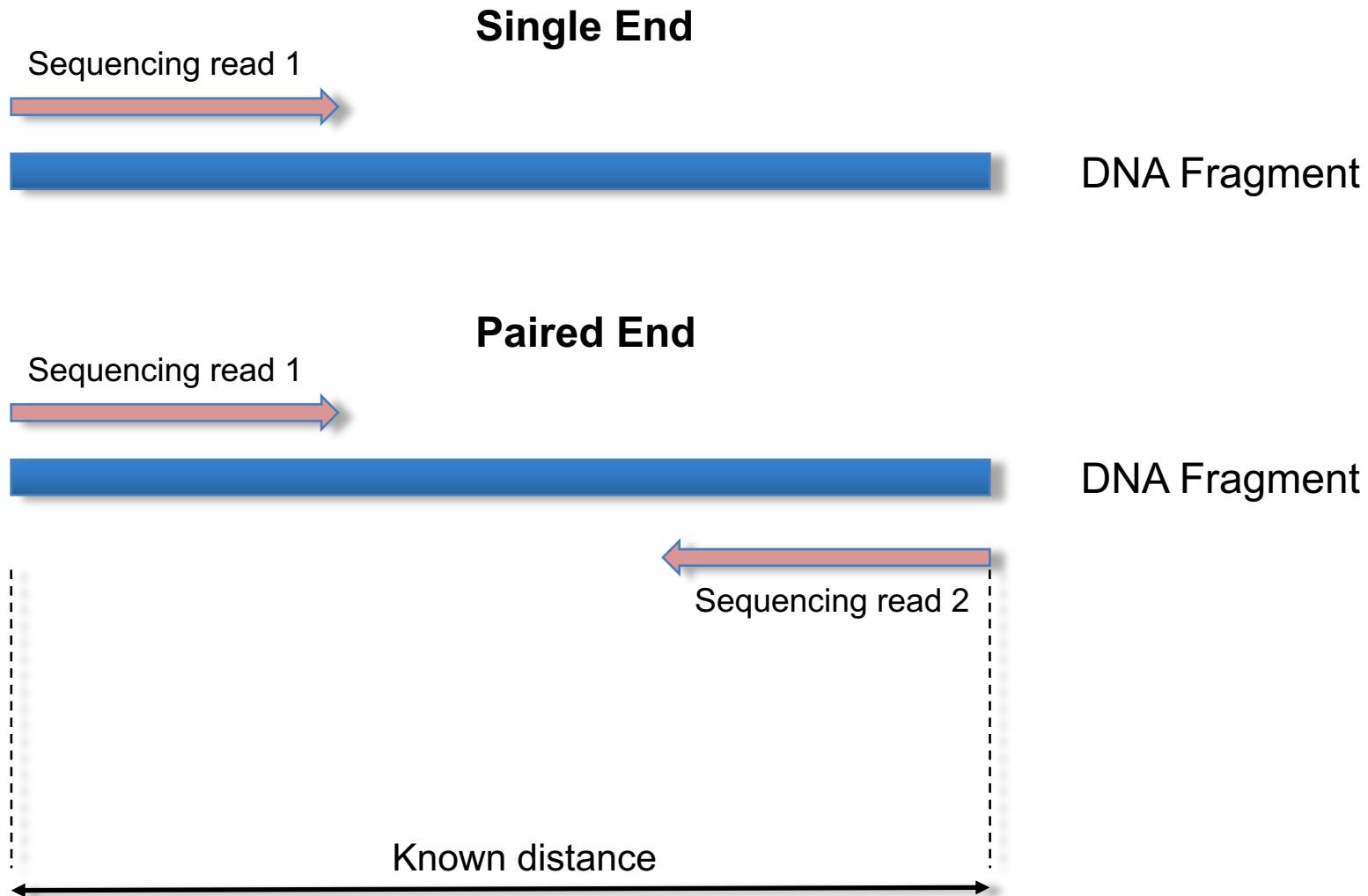
Classes of structural variations

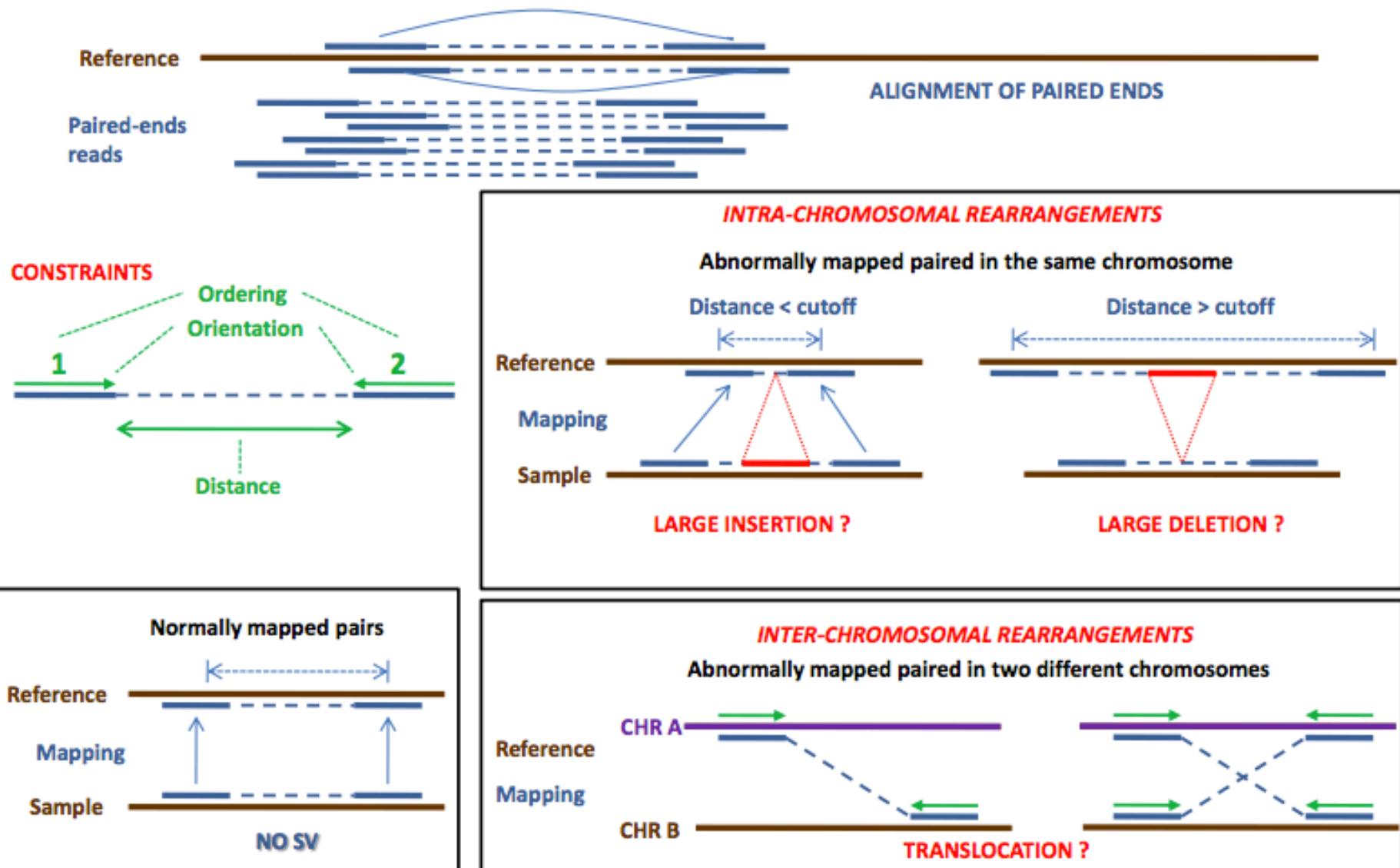
Alkan et al 2012



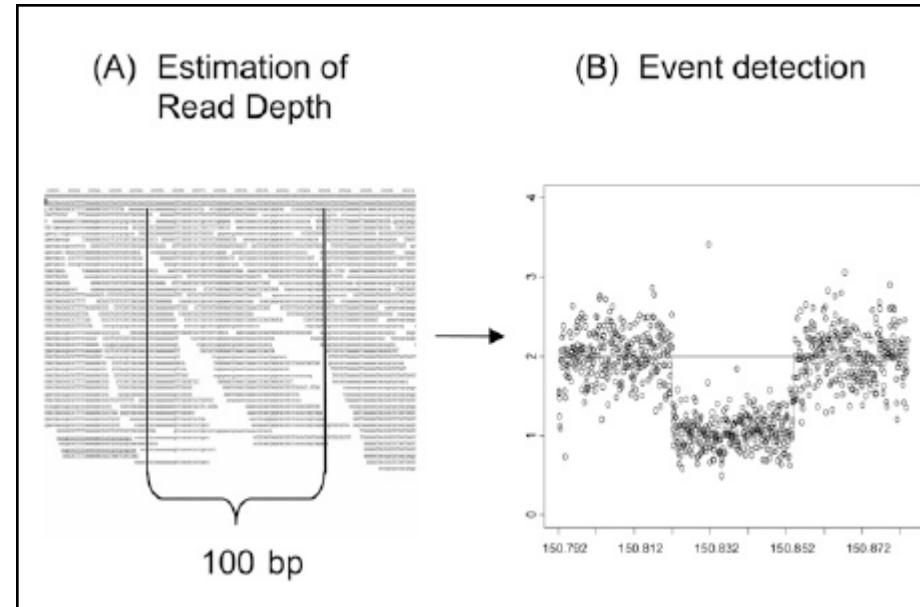
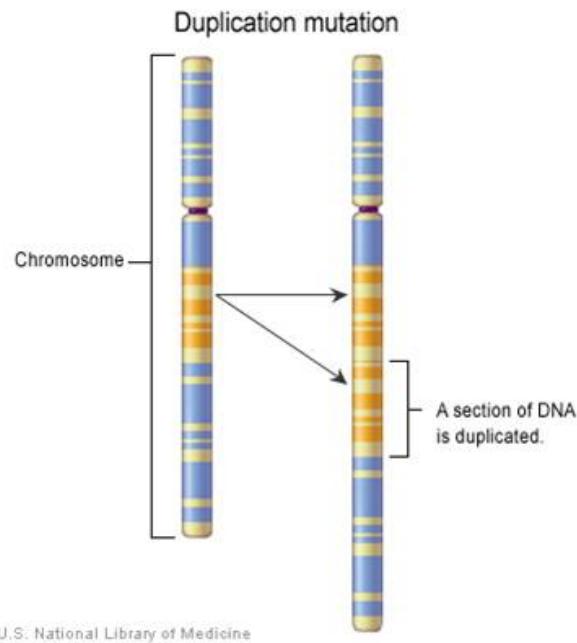
Intérêt du séquençage paired-end:

Résolution des repeats et des variants structuraux





Recherche de CNV (copy number variations)



- Attempts to infer variations in copy number from the **local read depth**.
- A strong GC% debiasing is required

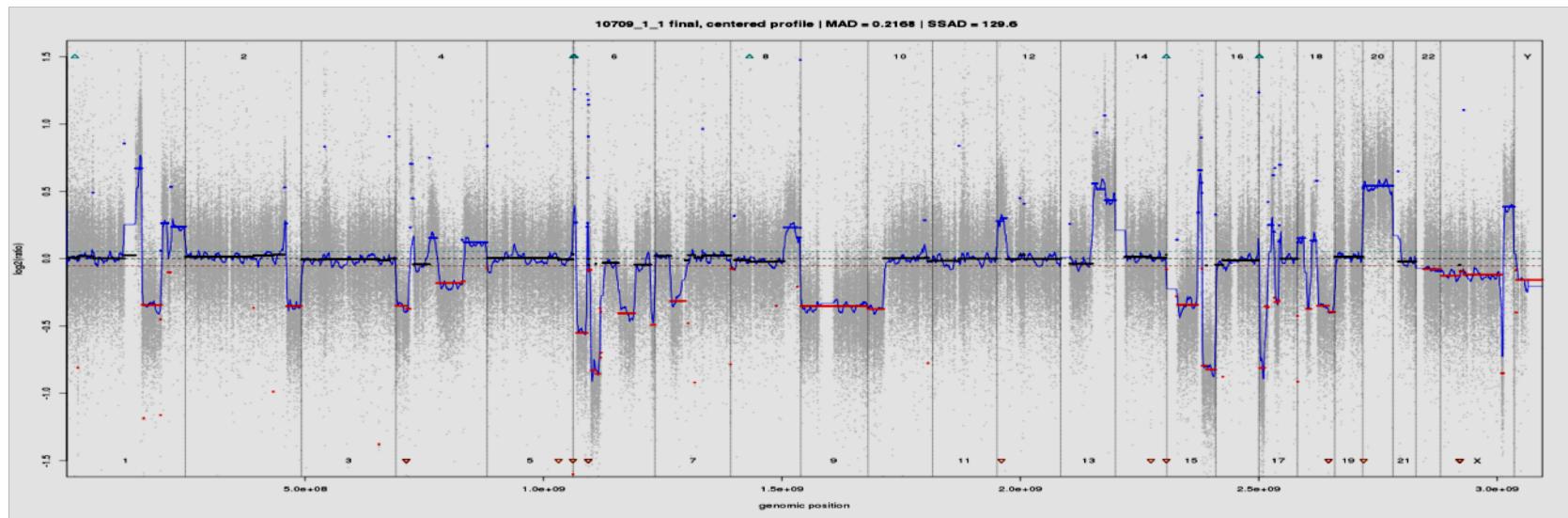
Voir cours Bastien Job

Yoon, 2009

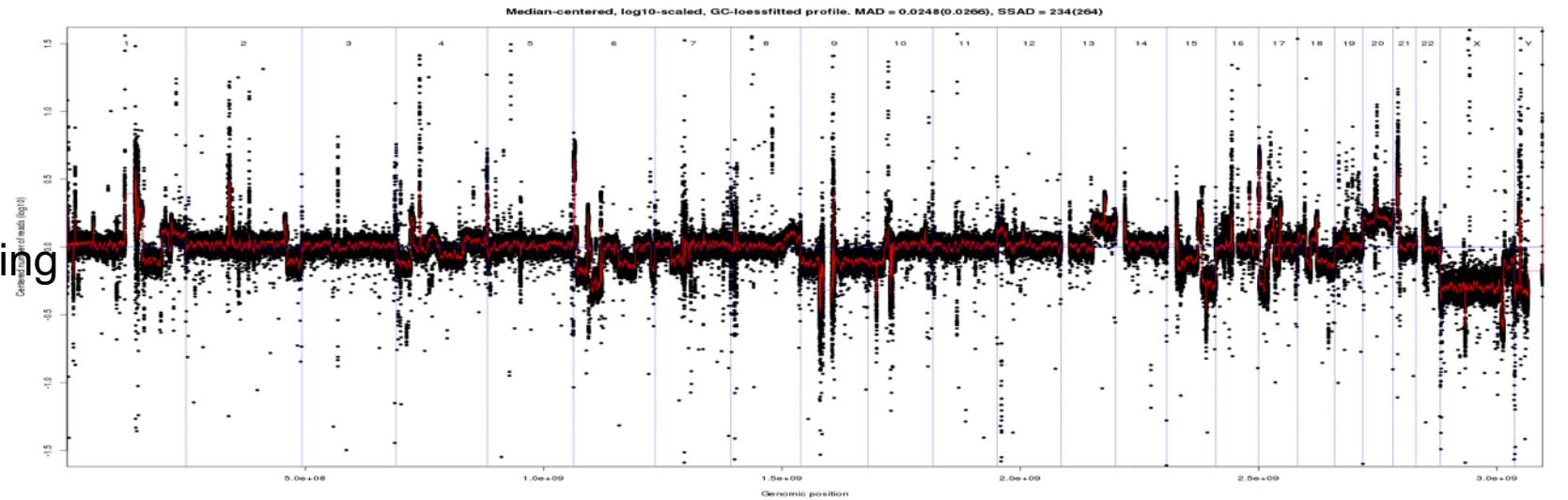
NGS vs CGH

CGH

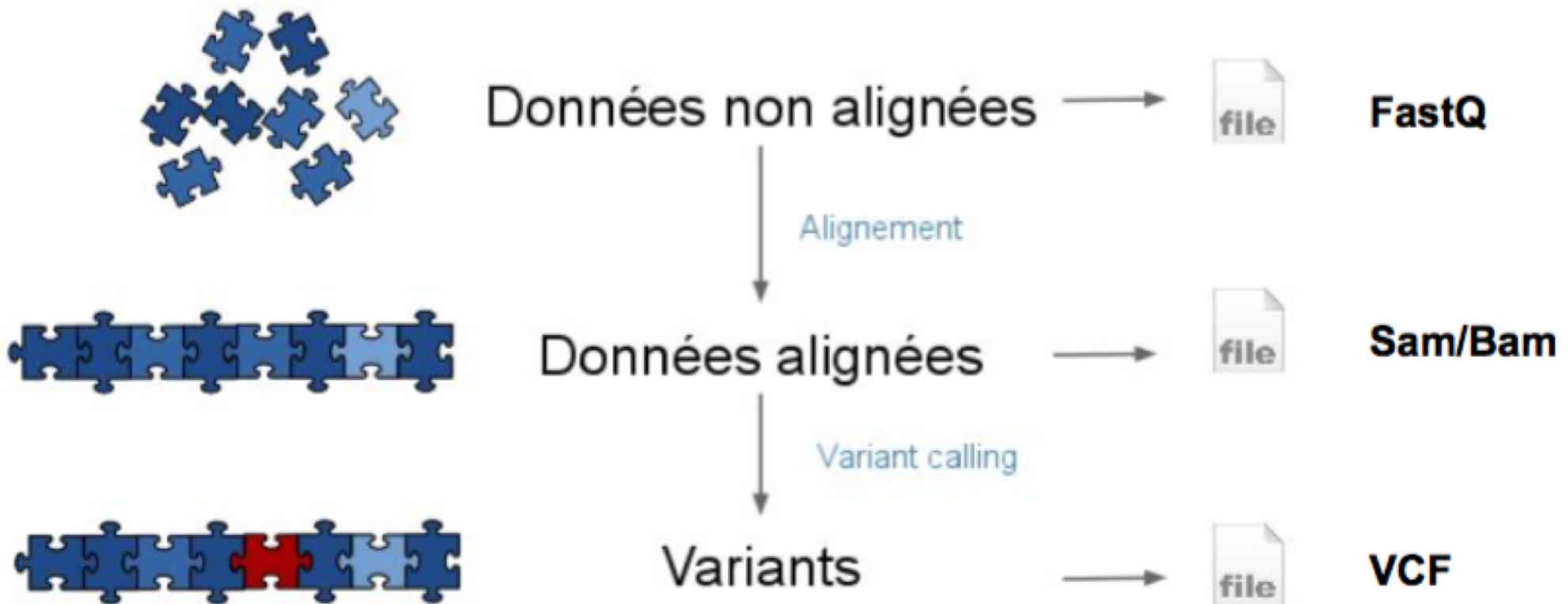
comparative
genome
hybridization



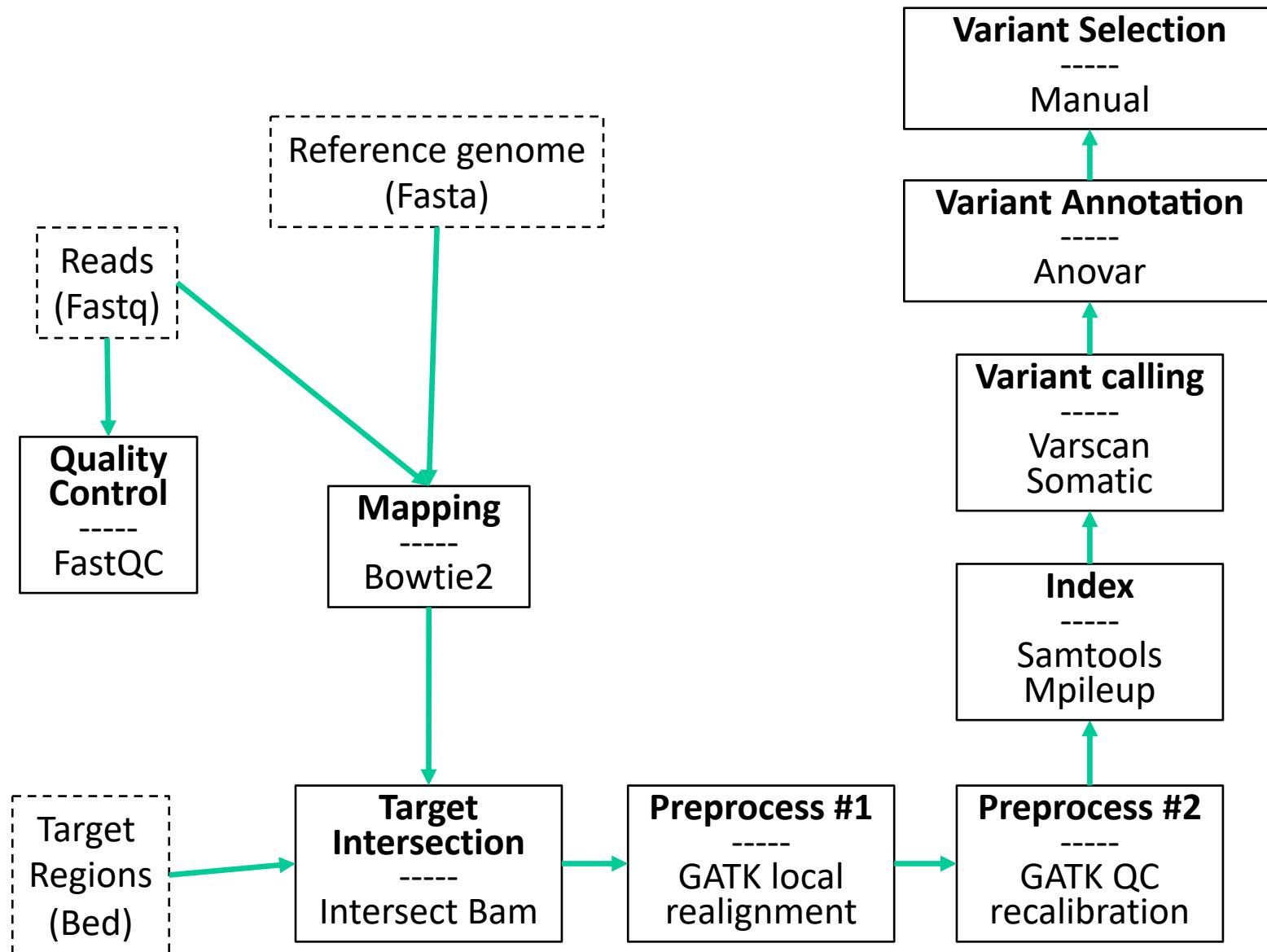
Sequencing



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



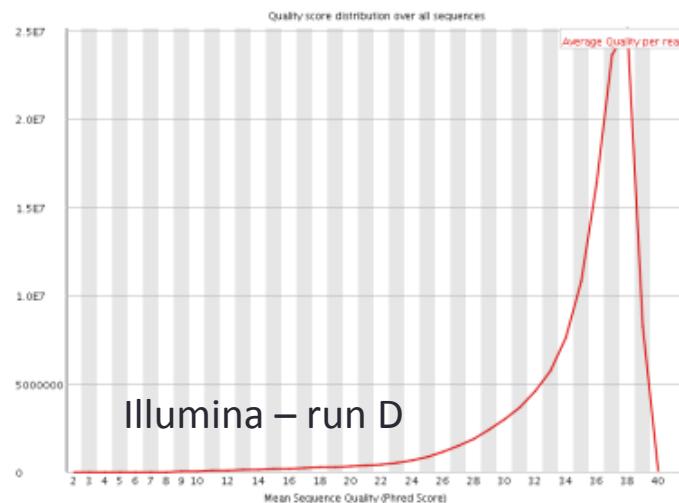
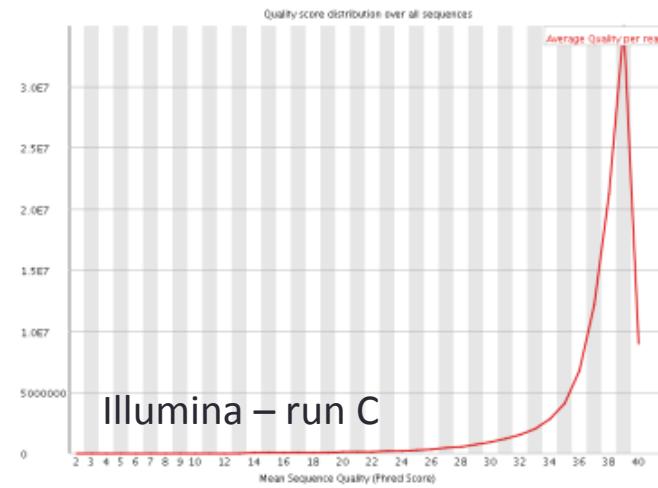
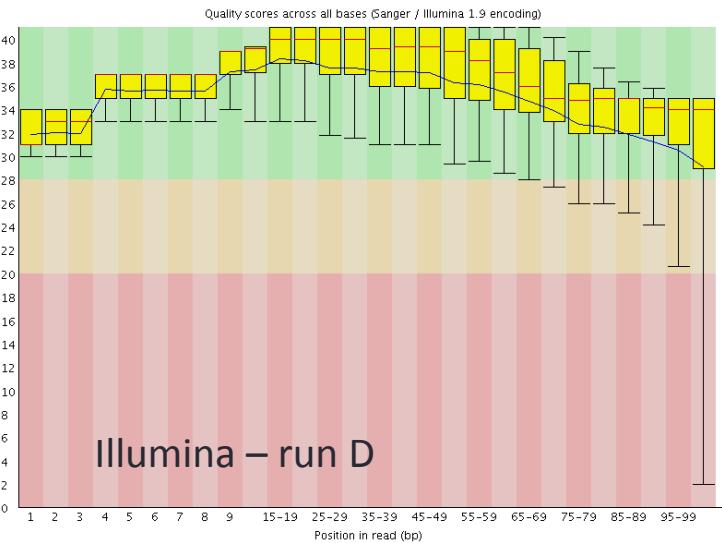
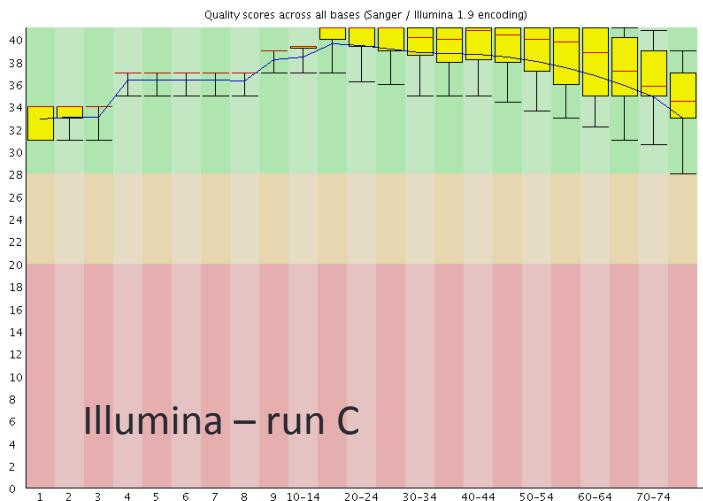
Solution:

Trim low quality bases from the 3' end of the reads

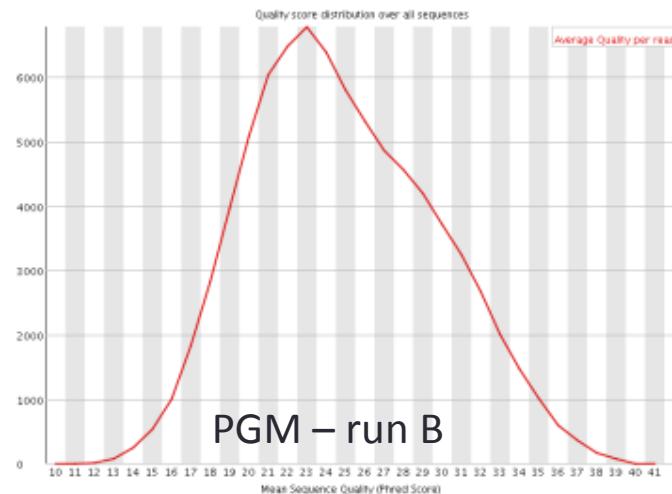
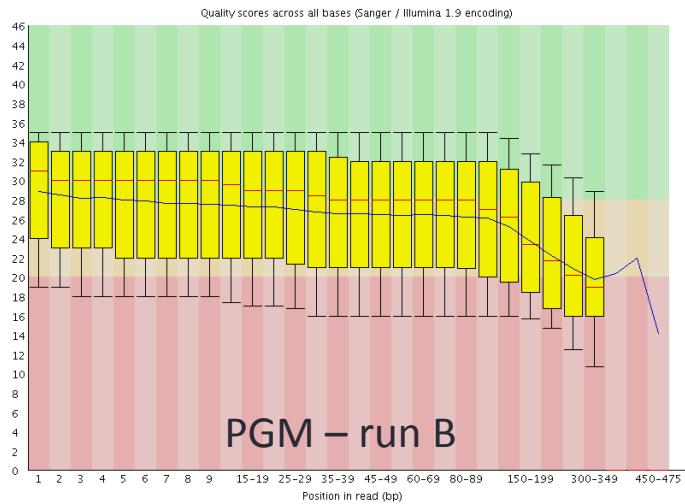
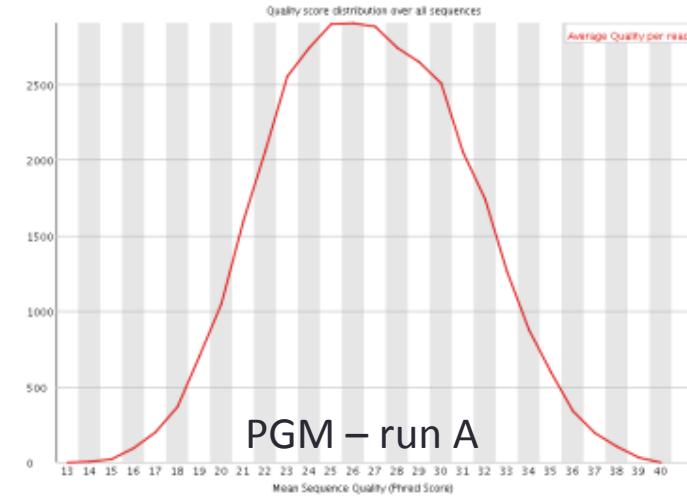
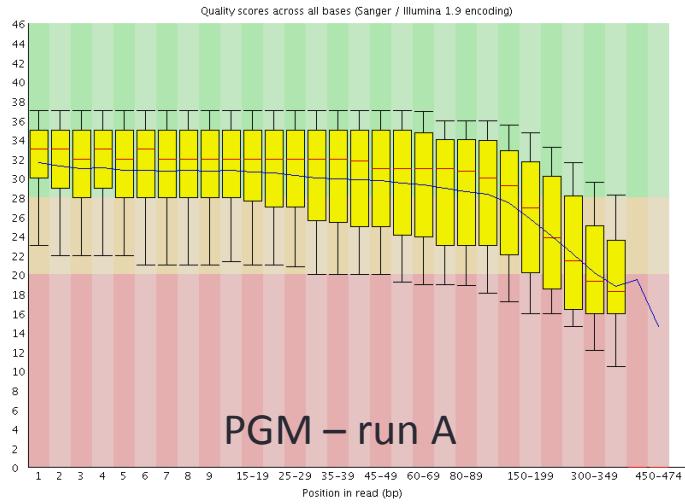
➤ Higher confidence in the sequenced information

(Trimmomatic)

Illumina

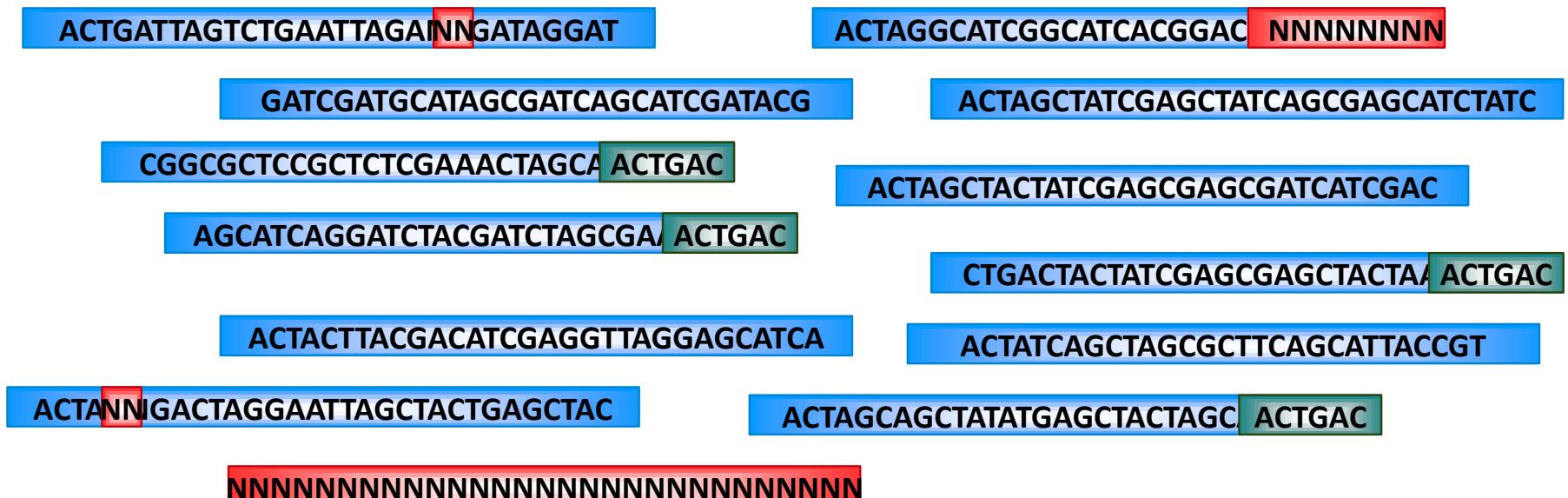


PGM

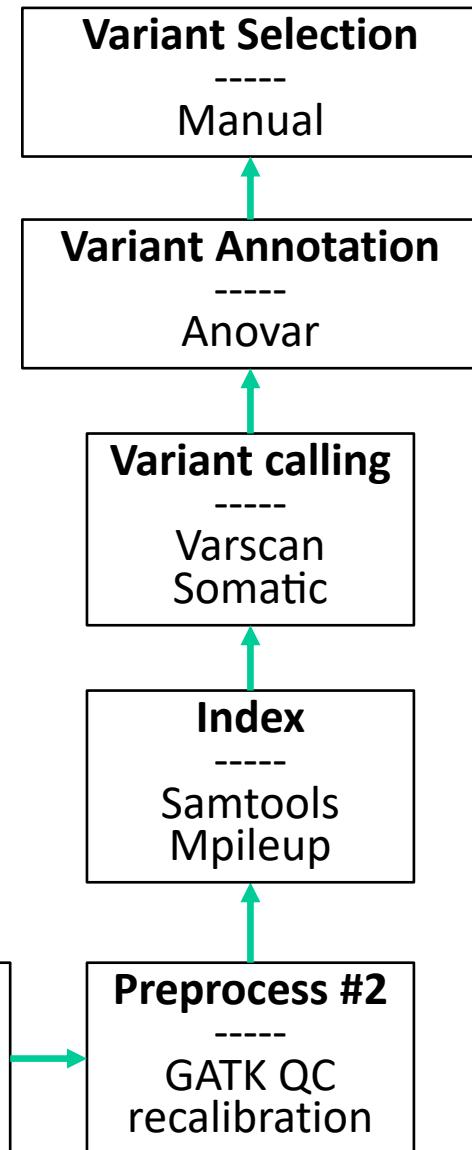
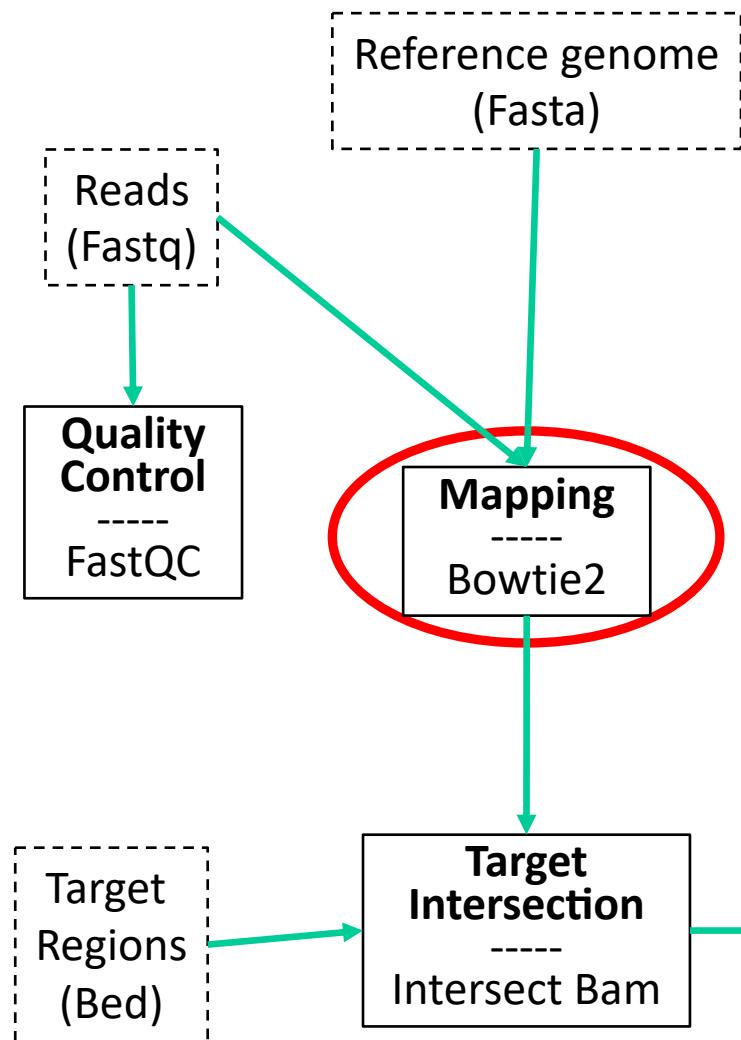


Trimming and discarding low quality reads...

A first Quality Control of raw reads is mandatory and can be established according to the application ('N', adapter sequences, barcode, contamination, etc.)



Processed reads: blue parts are to
be kept, green and red parts to be
removed



Alignment des reads sur le génome de référence



Algorithme en $O(mn)$

ACGTTACCGAATCGATCAAGTCGA
TAC



OK pour 1 read: $O(3.10^9 \times 100)$
Mais pour 1^8 reads???

« supercalifragilis-ticexpialidocious »

← →

Préfixe Suffixe

“GOOGOL”

Suffix array

Tableau trié de tous les suffixes
d'une chaîne de caractères

0 GOOGOL\$		6 \$	
1 OOGOL\$		3 GOL\$	
2 OGOL\$		0 GOOGOL\$	
3 GOL\$	→	5 L\$	→ (6,3,0,5,2,4,1)
4 OL\$		2 OGOL\$	
5 L\$		4 OL\$	
6 \$		1 OOGOL\$	

Propriété: toutes les occurrences d'une même chaîne
sont regroupées.

Suffix arrays

Exemple: trouver la chaîne **GO**

0 GOOGOL\$		6 \$
1 OOGOL\$		3 G O L\$
2 OGOL\$		0 GO OGOL\$
3 GOL\$	→	5 L\$
4 OL\$		2 OGOL\$
5 L\$		4 OL\$
6 \$		1 OOGOL\$

Most popular aligners for variant analysis

(support mismatched, gapped, paired-end alignment)

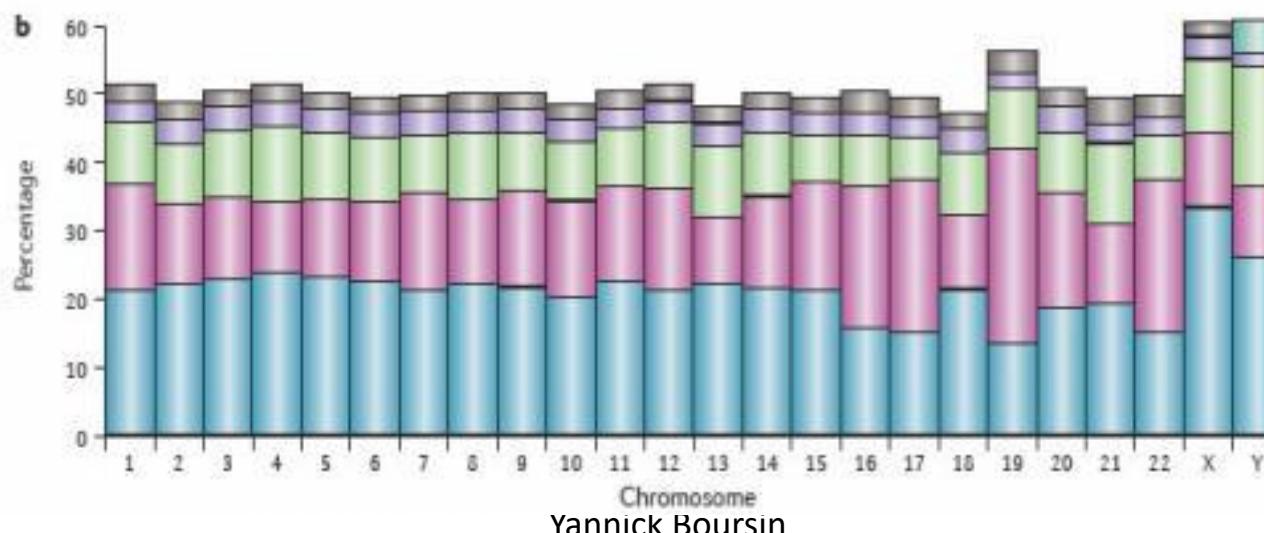
- BWA
 - Li H. and Durbin R. (2009)
- Bowtie2
 - Langmead B, Salzberg S (2012)

Alignment key parameters - Repeats

Approximately **50%** of the human genome is comprised of repeats

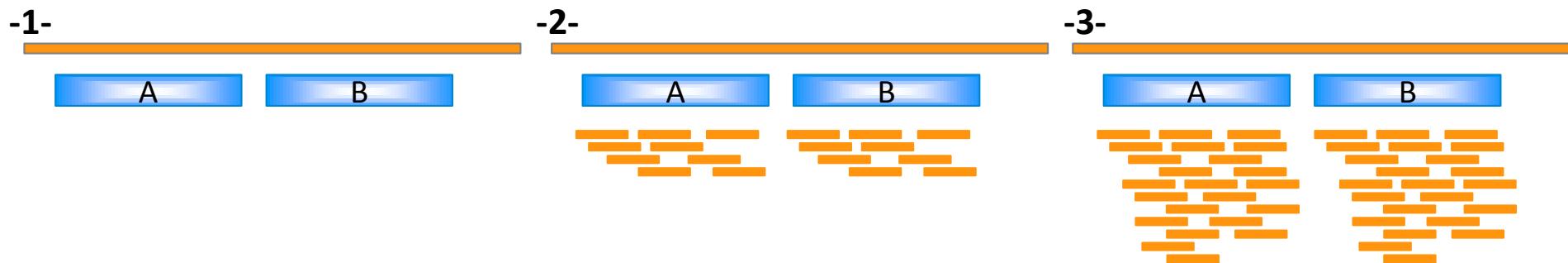
a

Repeat class	Repeat type	Number (hg19)	Cvg	Length (bp)
Minisatellite, microsatellite or satellite	Tandem	426,918	3%	2–100
SINE	Interspersed	1,797,575	15%	100–300
DNA transposon	Interspersed	463,776	3%	200–2,000
LTR retrotransposon	Interspersed	718,125	9%	200–5,000
LINE	Interspersed	1,506,845	21%	500–8,000
rDNA (16S, 18S, 5.8S and 28S)	Tandem	698	0.01%	2,000–43,000
Segmental duplications and other classes	Tandem or interspersed	2,270	0.20%	1,000–100,000



Alignment key parameters – Repeats – 3 strategies

- 1- Report only unique alignment
- 2- Report best alignments and randomly assign reads across equally good loci
- 3- Report all (best) alignments



Treangen T.J. and Salzberg S.L. 2012. Nature review Genetics 13, 36-46

Alignment key parameters – Using single or paired-end reads ?

The type of sequencing (i.e. single or paired-end reads) is often driven by the application.

Exemple : Finding large indels, genomic rearrangements, ...

However, in most cases, the pair information can improve mapping specificity

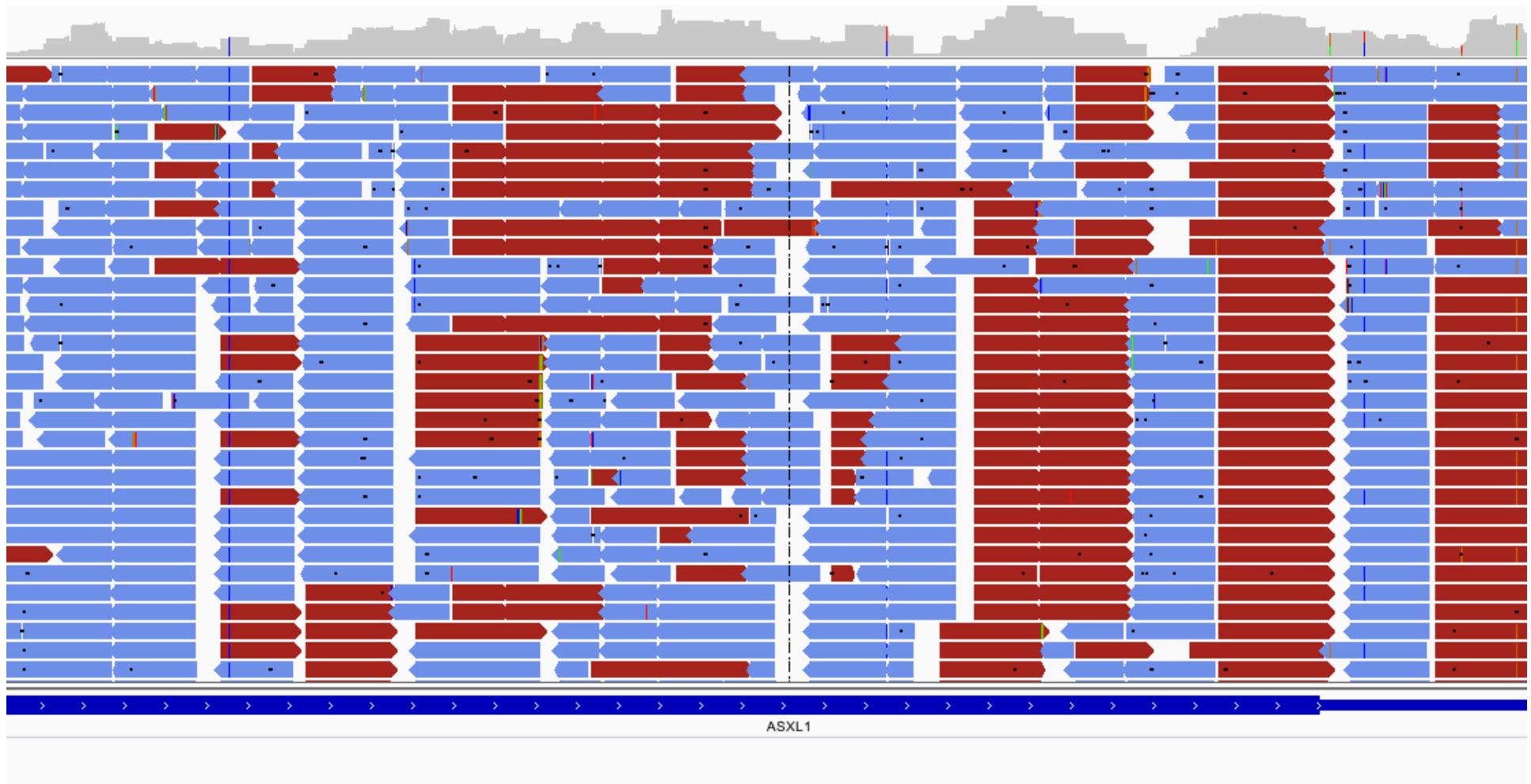
- Single-end alignment – repeated sequence



- Paired-end alignment – unique sequence



Reads alignés: le format BAM/SAM



Format BAM/SAM

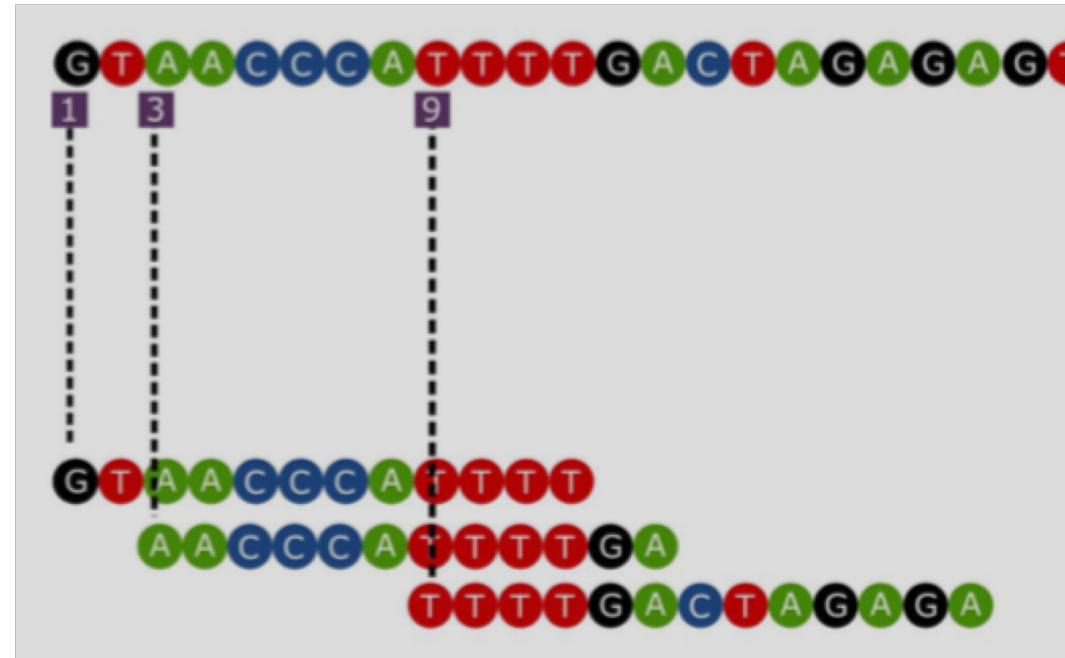
Contient les séquences alignées sur le génome

Dans l'idéal :

chr7 1324324 ACGTGCCTTCGCGT

chr8 1424324 GCGTGATGCGTAAG

chr8 1724354 GTATGTTATATGTA

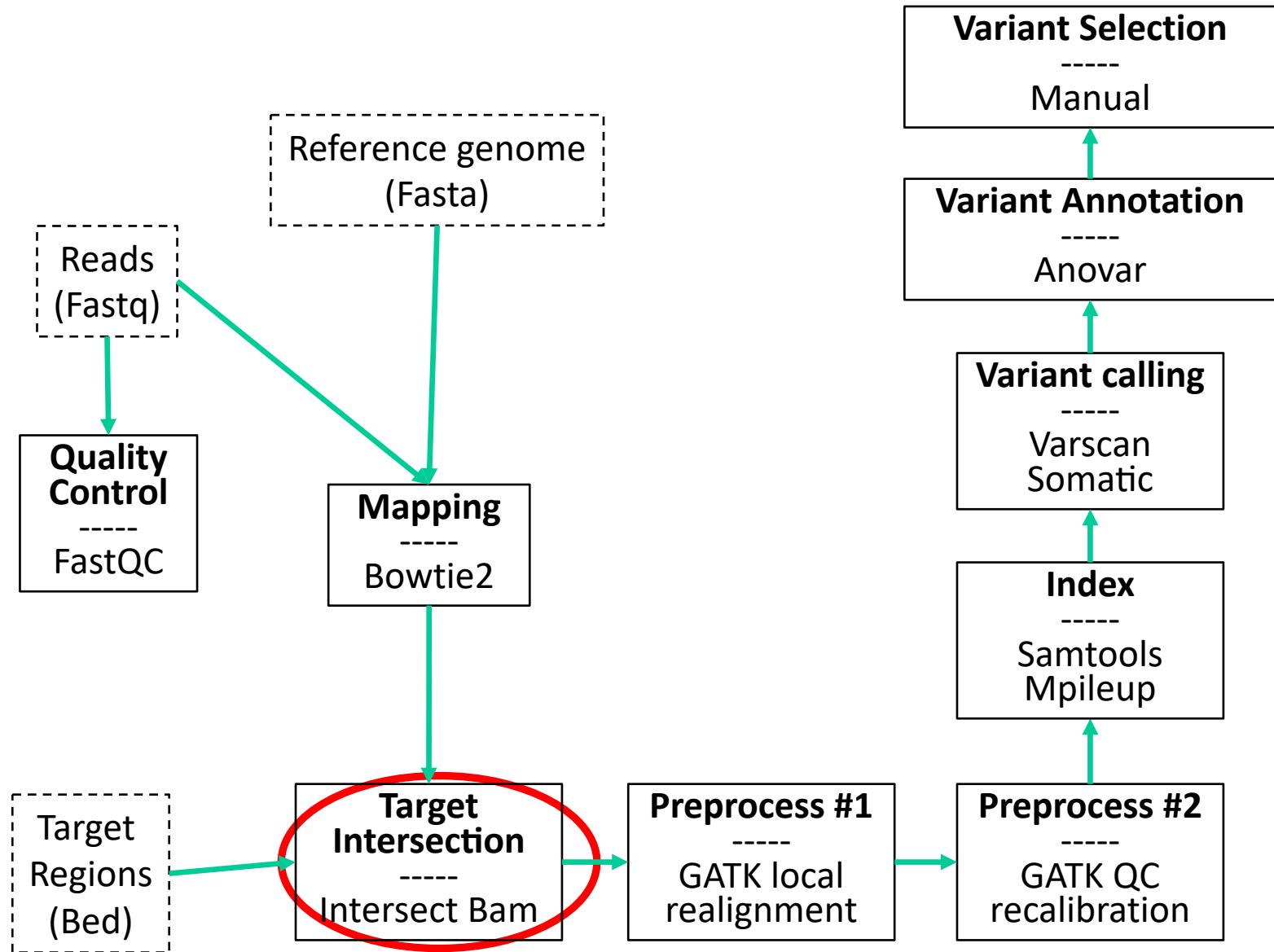


Format SAM/BAM

- Dans la réalité:

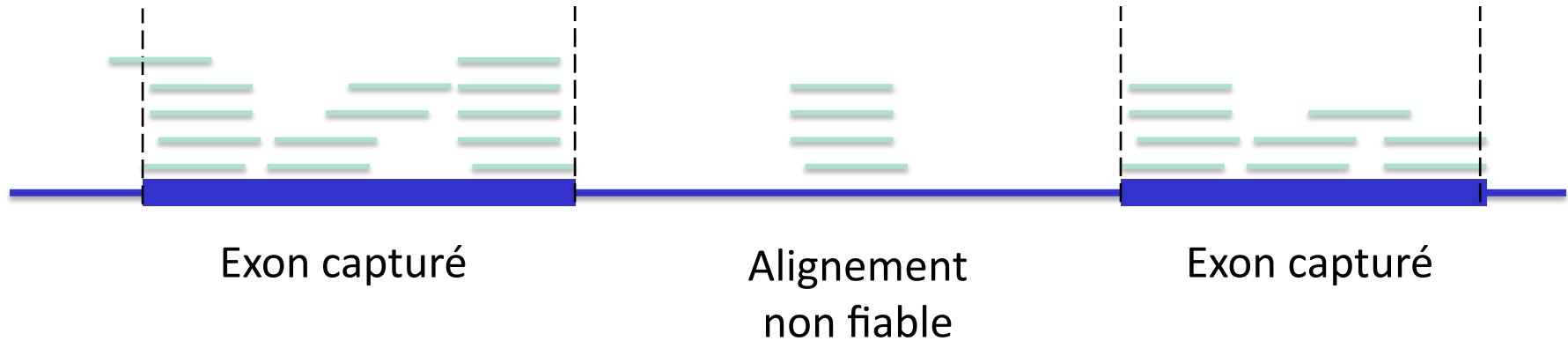
```
@RG  ID:group1  SM:1425_CD34  PL:ILLUMINA  LB:lib1 PU:unit1
@PG  ID:bwa  PN:bwa  VN:0.7.12-r1039 CL:bwa mem -M -t 2 -A 2 -E 1 -R @RG\tID:group1\tSM:1425_CD34\tPL:ILLUMINA\tLB:lib1\tPU:unit1 /root/myd
ERR166338.13782800  83  chr13  32890449  60  101M  =  32890343  -207  GGGACTGAATTAGAACAAATTTCAGCGCTT
ERR166338.13782800  163  chr13  32890343  60  75M  =  32890449  207  CACTAGCCACGTTCGAGTGCTTAATGTGGCTAGTGGC
ERR166338.26716588  99  chr13  32890406  60  101M  =  32890553  222  AATGTTCCCACCTCACAGTAAGCTGTTACCGTTCCAG
ERR166338.26716588  147  chr13  32890553  60  75M  =  32890406  -222  TTGCAGACTTACCAAGCATTGGAGGAATATCGTAAG
ERR166338.27259961  99  chr13  32890496  60  101M  =  32890558  137  ACCTCAGTCACATAATAAGGAATGCATCCCTGTGTAAG
ERR166338.27259961  147  chr13  32890558  60  75M  =  32890496  -137  GACTTATTTACCAAGCATTGGAGGAATATCGTAGGTAAG
ERR166338.63037998  99  chr13  32890496  60  101M  =  32890558  137  ACCTCAGTCACATAATAAGGAATGCATCCCTGTGTAAG
ERR166338.63037998  147  chr13  32890558  60  75M  =  32890496  -137  GACTTATTTACCAAGCATTGGAGGAATATCGTAGGTAAG
```

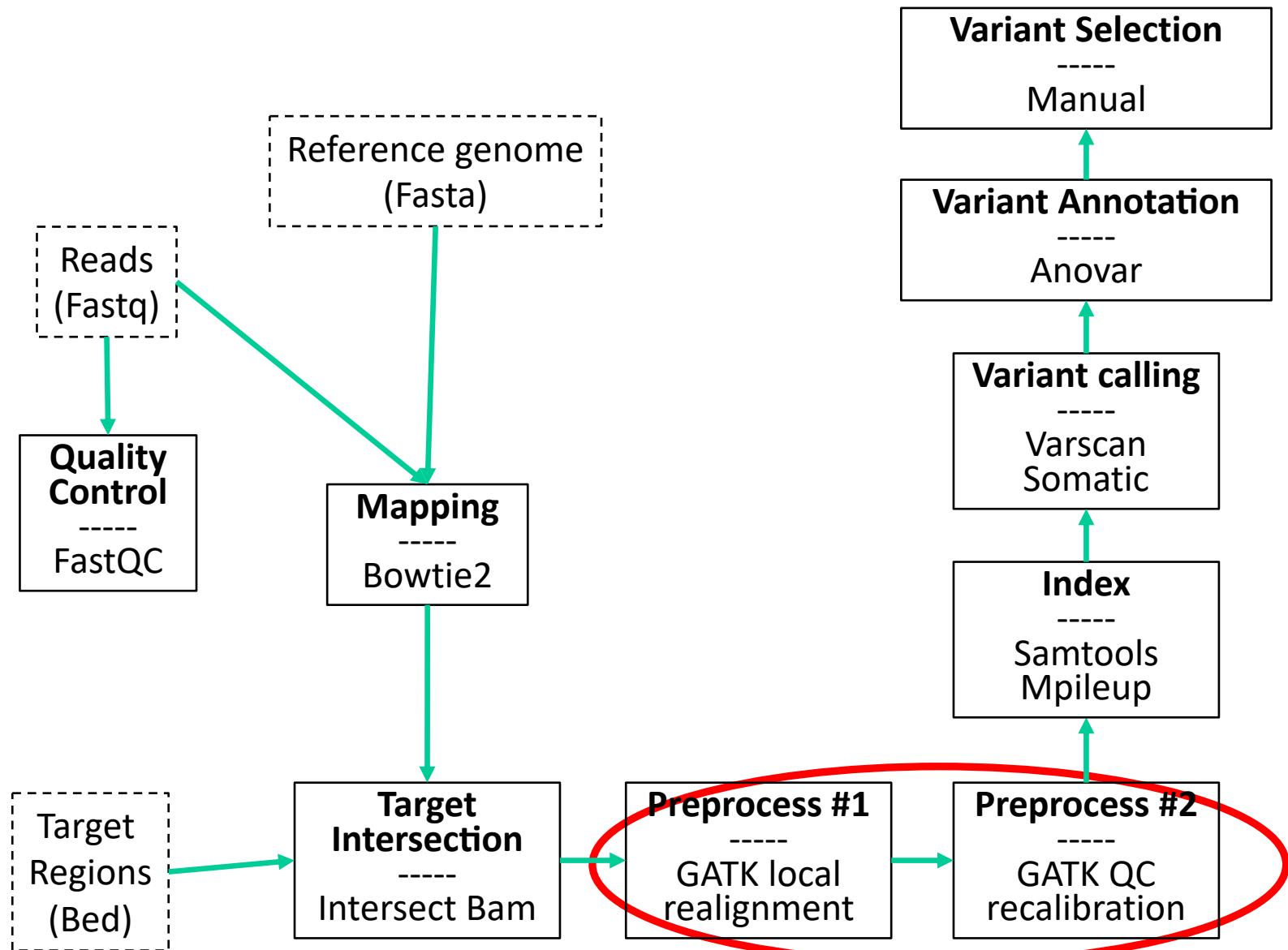
↑
read ID
↑
flag
↑
position
↑
CIGAR
↑
mapping qual.
↑
mate info



Target intersection

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





Why realign 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

Genome

CTACGAAGTAAAAAAAAGAGAGAGTTACT

CTACGAAGT - -AAAAAAAAGAGAGAGTTACT

CTACGAAGTAAAAAAAAGAGAGAG**TTACT**

Cost for 2 indels < 4 mismatches

CTACGAAGT - -AAAAAAAAGAGAGA

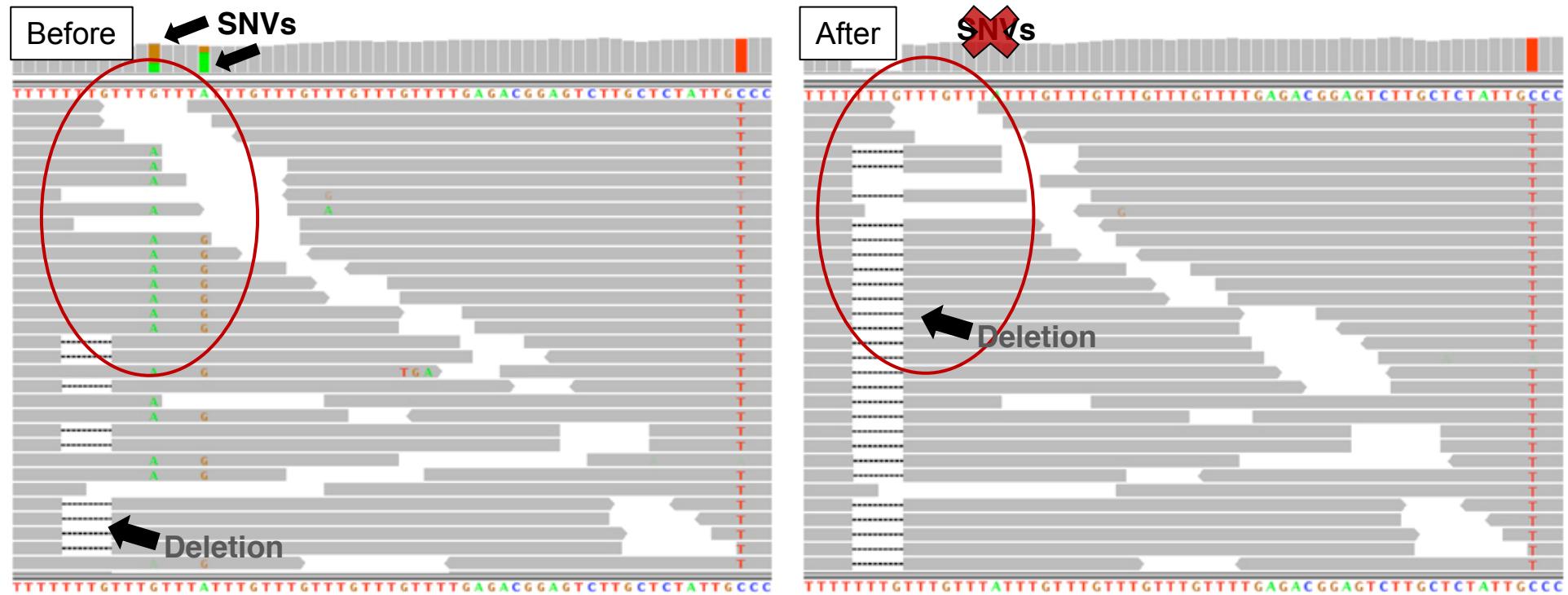
CTACGAAGTAAAAAAAAG**GAGAGA**

Cost for 2 indels > 1 mismatch

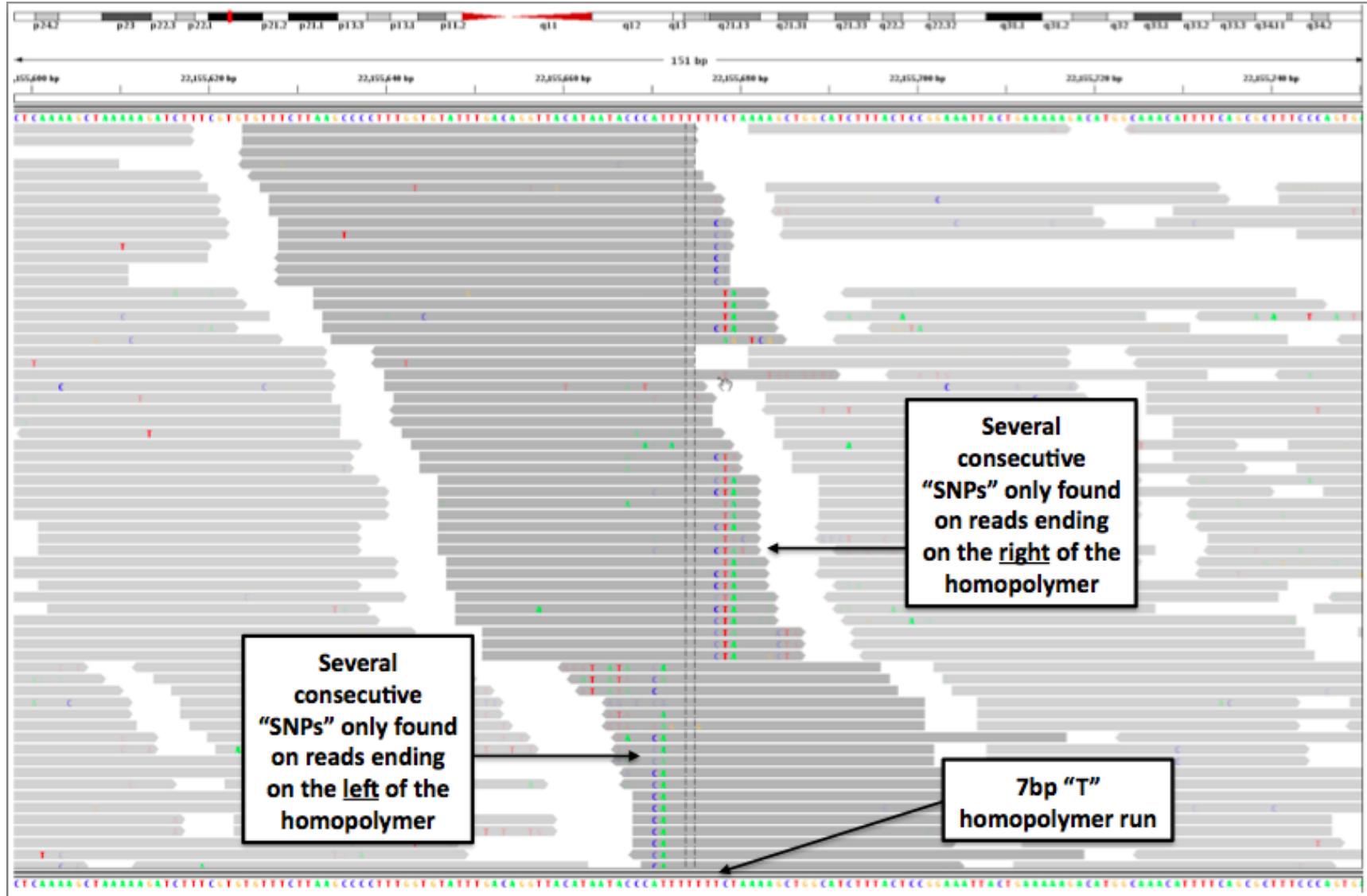
Read 1: 2 deletions

Read 2: 2 deletions

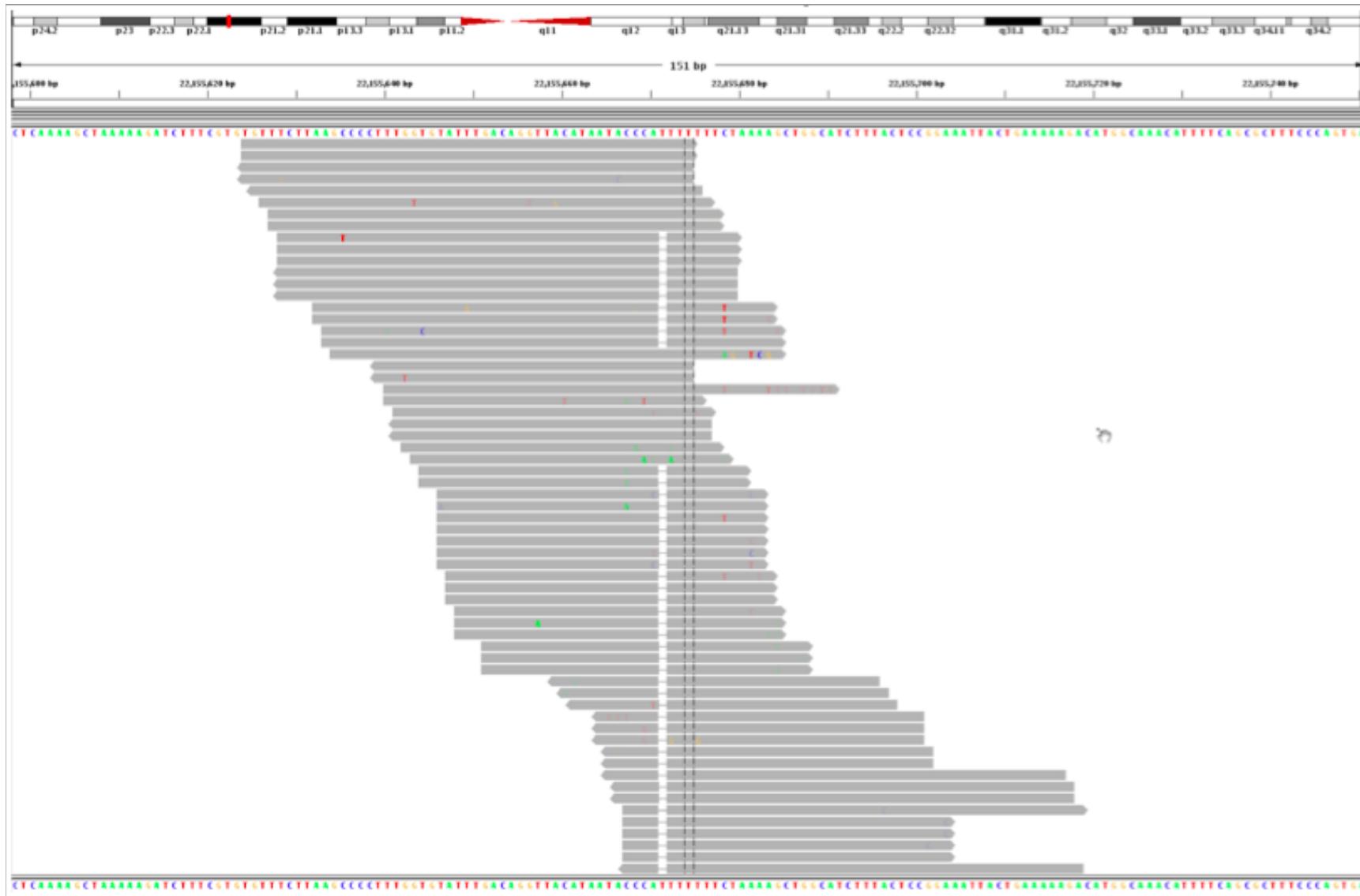
Local realignment around indels



Local realignment around indels



Local realignment around indels



Indel realignment in 2 steps

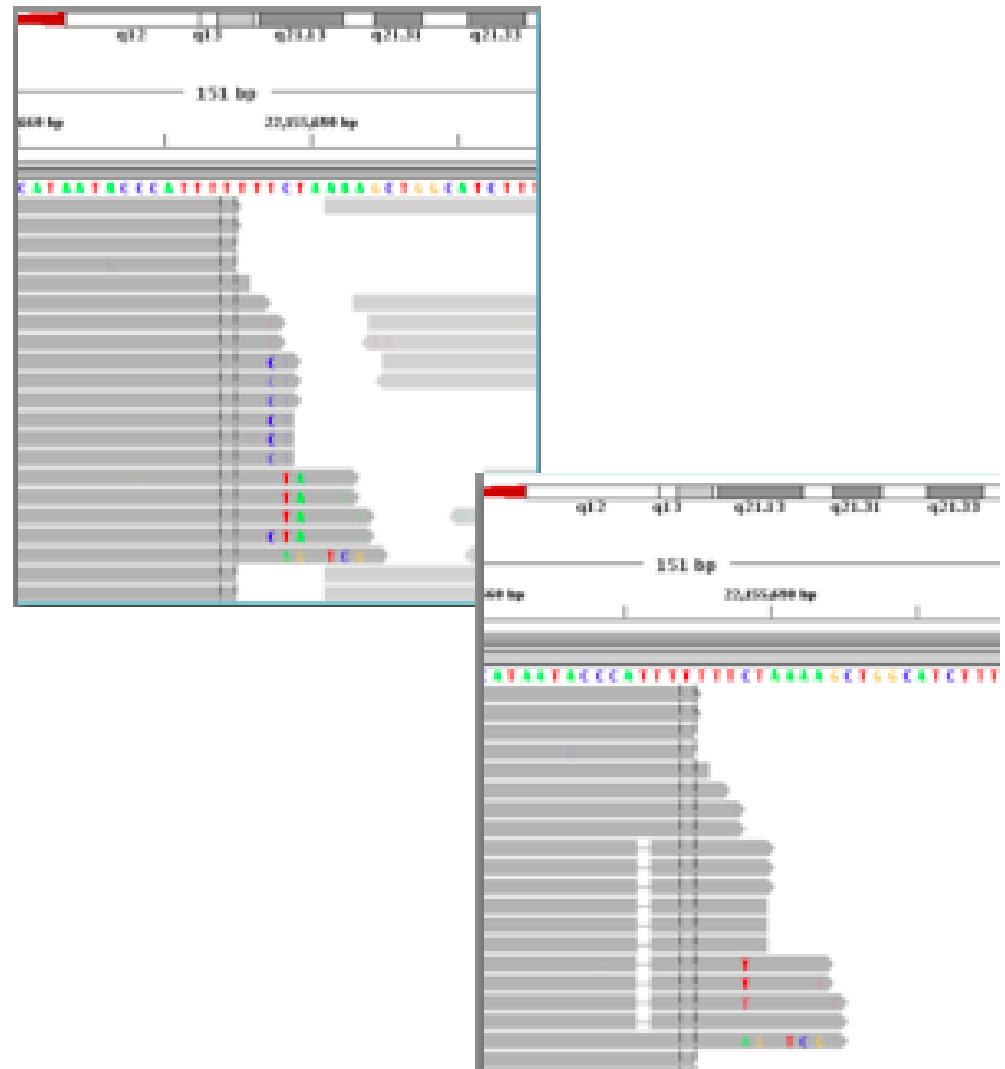
1. Identify what regions need to be realigned

- RealignerTargetCreator + known sites

Intervals
↓

2. Perform the actual realignment (BAM output)

- IndelRealigner

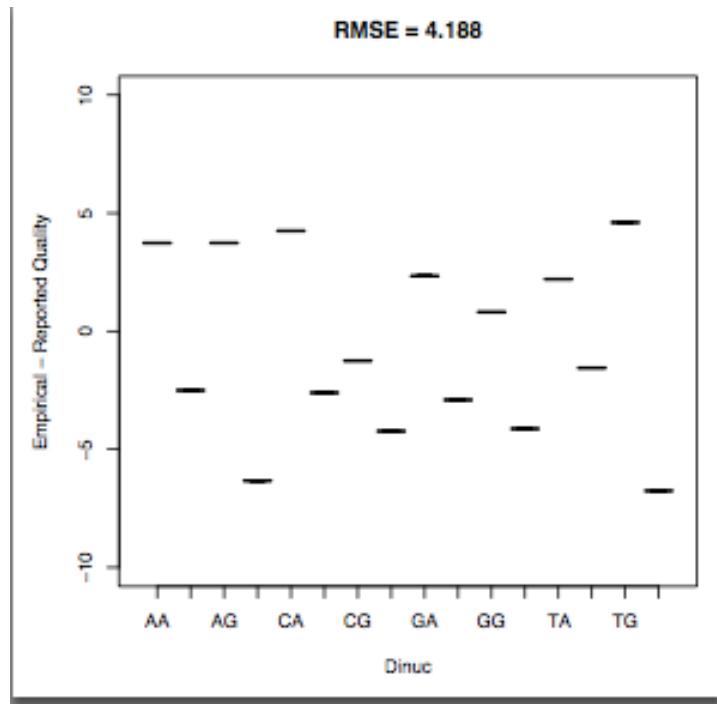


Types of realignment targets

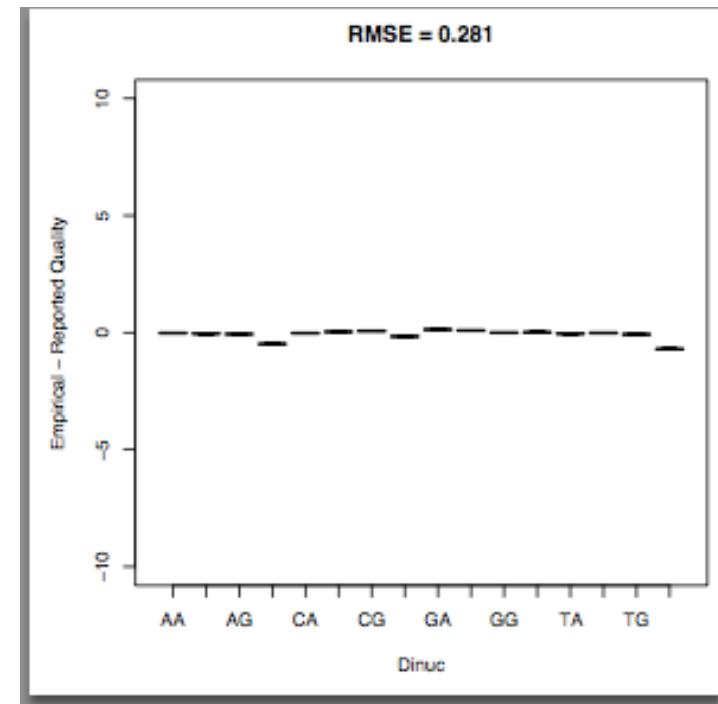
1. Indels seen in original alignments (in CIGAR, indicated by I for Insertion or D for Deletion)
2. Sites where evidences suggest a hidden indel (SNV abundance)
3. Known sites:
 - Common polymorphisms: dbSNP, 1000Genomes

The quality scores issued by sequencers are biased

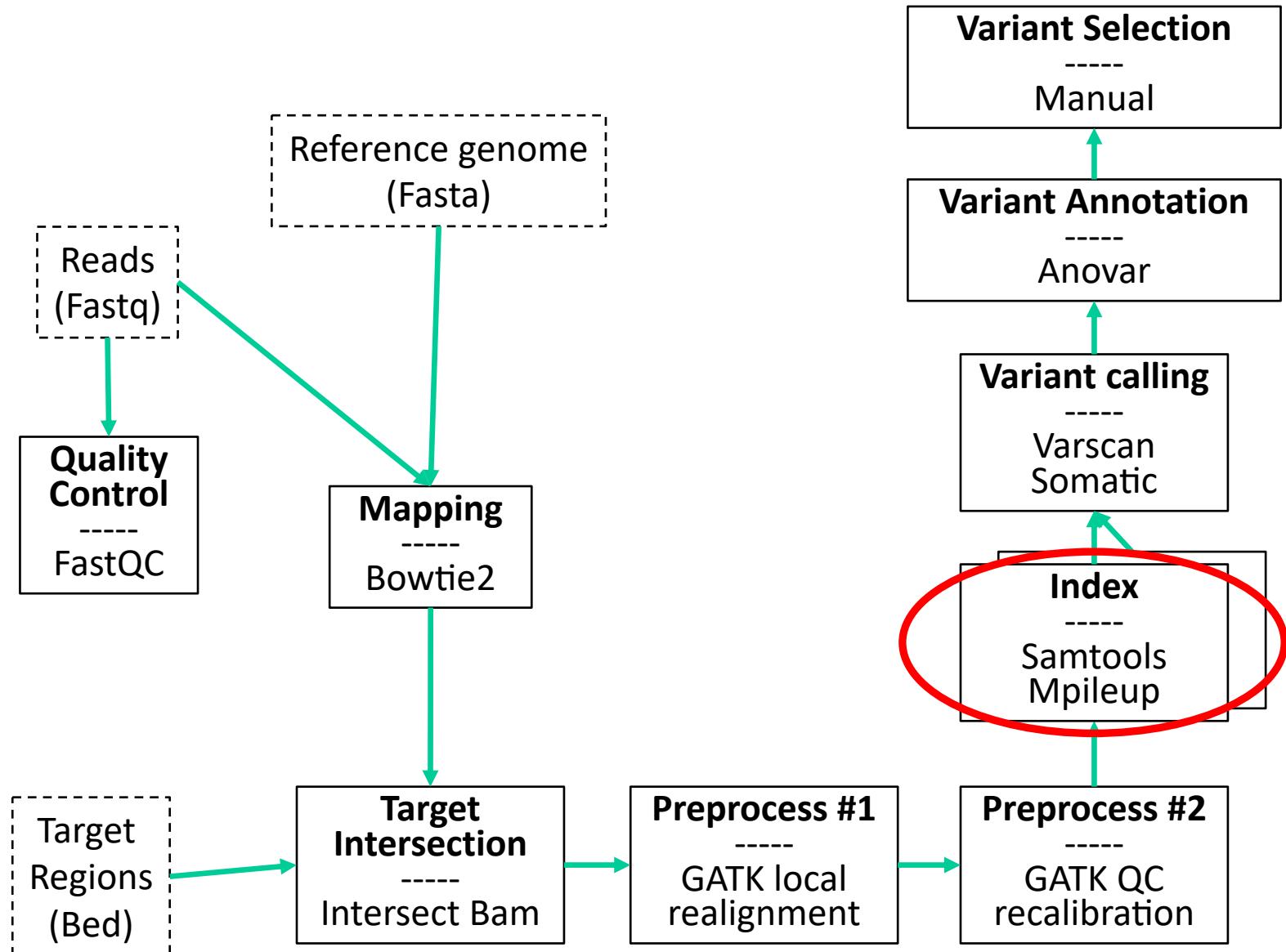
- Quality scores are critical for all downstream analysis
- Systematic biases are a major contributor to bad calls
- Example of sequence context bias in the reported qualities:



before



after



PileUp: Pourquoi un nouveau format de fichier?

Rappel BAM:

```
@RG  ID:group1  SM:1425_CD34  PL:ILLUMINA  LB:lib1 PU:unit1
@PG  ID:bwa  PN:bwa  VN:0.7.12-r1039 CL:bwa mem -M -t 2 -A 2 -E 1 -R @RG\tID:group1\tSM:1425_CD34\tPL:ILLUMINA\tLB:lib1\tPU:unit1 /root/myd
ERR166338.13782800  83  chr13  32890449  60  101M  =  32890343  -207  GGGACTGAATTAGAACAAATTTCAGCGCTT
ERR166338.13782800  163  chr13  32890343  60  75M  =  32890449  207  CACTAGCCACGTTCGAGTGCTTAATGTGGCTAGTGGC
ERR166338.26716588  99  chr13  32890406  60  101M  =  32890553  222  AATGTTCCCACCTCACAGTAAGCTGTTACCGTTCCAG
ERR166338.26716588  147  chr13  32890553  60  75M  =  32890406  -222  TTGCAGACTTACCAAGCATTGGAGGAATATCGTAAG
ERR166338.27259961  99  chr13  32890496  60  101M  =  32890558  137  ACCTCAGTCACATAATAAGGAATGCATCCCTGTGTAAG
ERR166338.27259961  147  chr13  32890558  60  75M  =  32890496  -137  GACTTACCAAGCATTGGAGGAATATCGTAGGTAAG
ERR166338.63037998  99  chr13  32890496  60  101M  =  32890558  137  ACCTCAGTCACATAATAAGGAATGCATCCCTGTGTAAG
ERR166338.63037998  147  chr13  32890558  60  75M  =  32890496  -137  GACTTACCAAGCATTGGAGGAATATCGTAGGTAAG
```

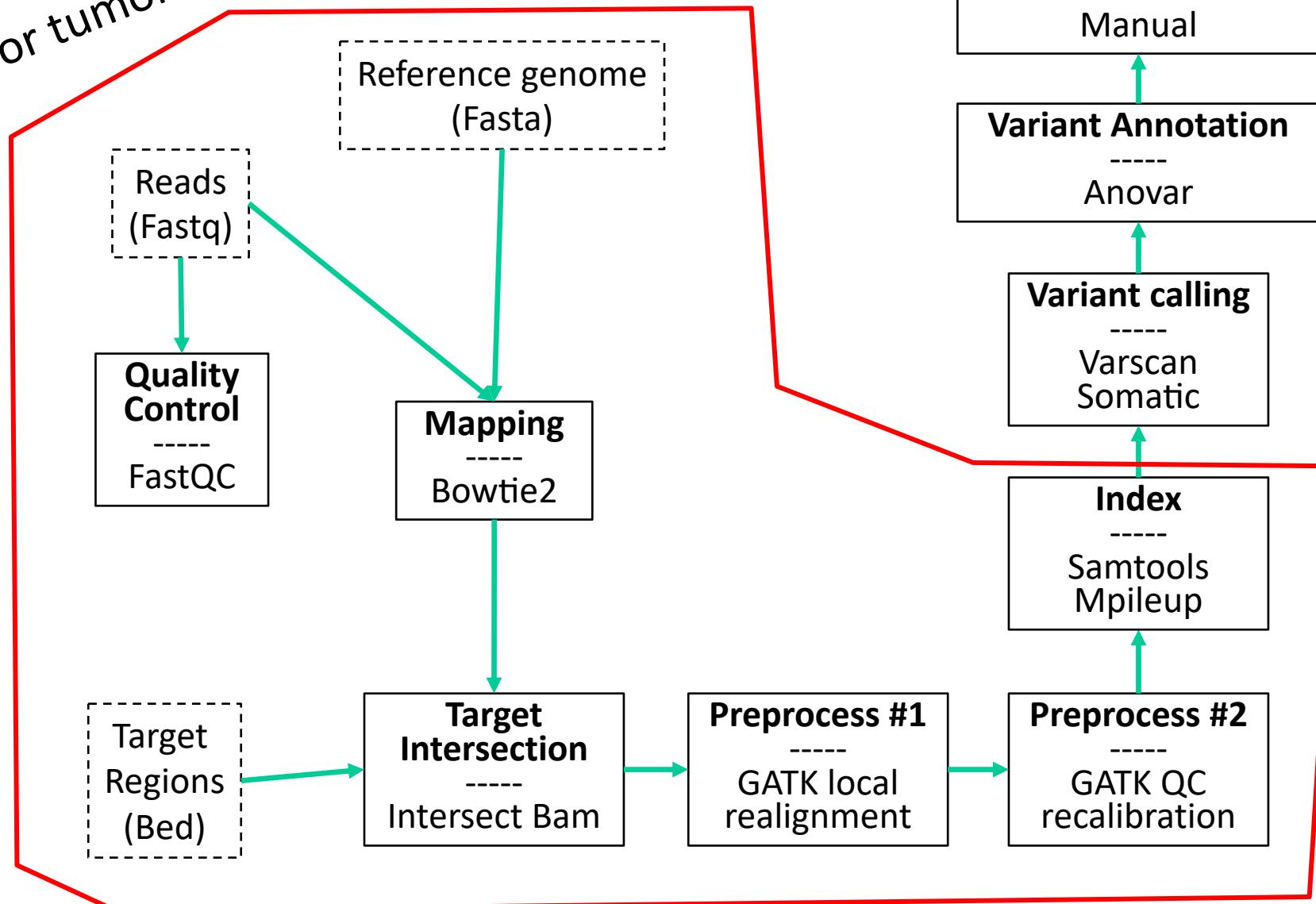
Pileup format

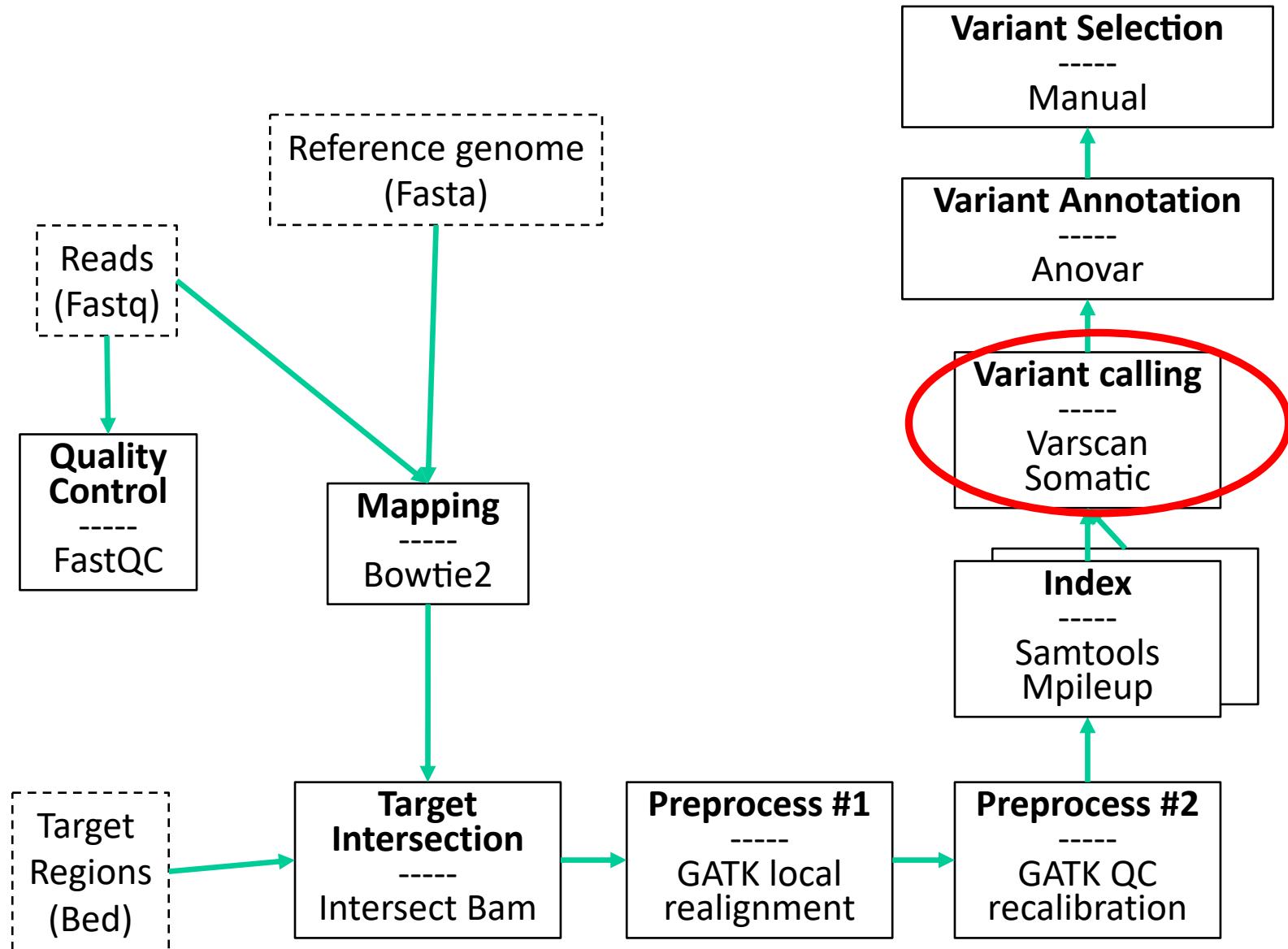
```
|chr11 26692809 T 65 ,,$,$.....@C@cFFCFHHGHIG9JIJJDA?Ge@fJHGJmIFIj$  
chr11 26692810 A 62 .,$,.....^]. hFFCHGHHHFg<JIJGDCHJdEeJJFJKIFHljJq$  
chr11 26692811 C 62 ,gGgg,gg,gggggg,G,g,G,.g,gg,g..gGgg,ggg,,g,ggg,,g..gg,G CC>HHHHHHG?JIIJDCGI`@eJJIIJJEElkIpI?[J:IBJ$  
chr11 26692812 A 64 .....,CC@DGHHHGD<JJJlJ@CGJaIgJJIJlJDIjIjq$  
chr11 26692813 A 62 ,$,..... CCCDFHHHHH>HHHGJAFEJbEFIJIJkJEHm_IoS  
chr11 26692814 A 62 .....,^]. CDFFFFHGCHHHHHCCHIUBfJJJnJEDkfInHF$  
chr11 26692815 G 61 .$. C=FFFFFH:HHHHHC?EHOGeJJIKICmhnGbJII1GBII$  
chr11 26692816 A 62 .....,BDFFFHD2HFHHHDADHZBgIJIHkGEFmgHmAFChIDG1DBI$  
chr11 26692817 C 62 .....,?FFFFDD=FHHHDHC<H*GfJJJJjIDCkgJoCA2fIDI9IEH$  
chr11 26692818 A 60 .....,?BDDFD?DAFC=?H]EaHJHIIHDEjfHmAF[I?GHIJHJ9$  
chr11 26692819 C 63 .$. ?@CCCFDBFFFEDDFHhIgJJJmJDJmjJn999$  
chr11 26692820 A 62 .$. @C@CCD?FFFDFD?BHiGiJGJmJDIliJnDH<kJFH:IDJJ$  
chr11 26692821 T 61 ,$,,$. C@CCFDFFFBFBCADFhHfHJJjICHlhInEE2iJ$  
chr11 26692822 A 58 ,$. BC?FFF:FCDDFhHfHJJjJCClhJn9A<jJEJEI:JJFII$  
chr11 26692823 T 57 .....,C;CCCCFEDDFcFfHJJjkJDHnjJnCH<gJEFJ<IJGEHH$  
chr11 26692824 T 57 ,$,,$. C8C@C@CCBDDg8hHHJkIEBmjJm9E<fJ>G?JDHIJIFI$  
chr11 26692825 A 55 ,$,,$. B@@CC@DFdFjHHHJkJFFmlJ@HGHijDD@JGIJJHJ@JJ$  
chr11 26692826 G 52 ,$,,$. @BA?@B]DiFHHDhjI@GnJnHF@fJCJ@JFHJJGJBIIJJHH?F:EEFjh$  
chr11 26692827 C 50 .$. 9??@UBiFHFhiHCEmjJgIBAeI<JBJADJJHJGJJJHHEC>FEFcJC  
chr11 26692828 A 48 ,$,,$. ?@^DiFHGHiHDImjJlH@<fJ9JHJDFIJEJJJJJHHEE;FCFnJ@  
chr11 26692829 A 46 .....,g@IFFHHihDHLkJmCFIiJ<JFJAICJiJFJiJJHEEH;DDFnJ>  
chr11 26692830 G 46 .....,G;hFFFHiHDDkkJl4HAcJ3JBjHJHJiJJJJGFf?ADEmJC  
chr11 26692831 A 46 .$. I?kCFDFFGDHlmJmFG9hJFJ@JDIEJHIIJJJGHFHEFAEmJE  
chr11 26692832 T 46 .....,^], eCFDFFFDHmmJj<FIgJAj@JFIEJEIGIJJJJJEHHAECGmJED  
chr11 26692833 A 46 .$. ,$,.....^], hCFDFFFDGmmJkCEHijJJEBFGJHICIJJJJIGC=E=HmJCDD  
chr11 26692834 T 43 .....,CDFFDFmnHm3AGiJ9I:IBIGJEJDJJJJHGG=HHHmJEBD  
chr11 26692835 C 45 .....,CCFF@DmmGLFE<gJ>ICJ>JFJJJHJiJJGGIE3EHGHmJCAD  
chr11 26692836 C 45 .$. CBCFBnnHk<CEbJCJ?IEJiJFIGJGJJJDID@CCIHNJ@CD  
chr11 26692837 T 44 .$. @CCDFmmHmD>DhJ<I:FBJJJiJFJJJJ>GE=:I@IJmJCCC  
chr11 26692838 T 42 .$. CCFDmmHm4FAeJ:H?JGJJiJIIIIJiJJCIGGHDJinJFDC  
chr11 26692839 G 42 .$. CD@llFmF?AkJFJ?JHJJJiJJJJJJGGICEEGJFnJFDD  
chr11 26692840 A 41 .$. CBmlFm>A?kJBJ@IJJJJGJiJJJCIGCHAHH@JmJDEE  
chr11 26692841 T 40 .$. @mlFmDFCKJCGIGHIIIGJGJJJHGDCCHFCEGJiLJ@DD  
chr11 26692842 G 39 .$. lkFjAD=jI2IAGFHGIHIFJiHIJGf@=FIIGnJADD  
chr11 26692843 T 39 .....,UDkDD=hH2HAEHAFGHCHFIHFIICGF7@FHGFmI=CC  
chr11 26692844 A 36 .....,j@iDBdFG2FFBEEGECIGDII=FG7.BFICKIAAC  
chr11 26692845 C 39 .....,MC]==4iH2J3JJJEJGJGHJJJJHHJHBCCJJoJHEE  
chr11 26692846 T 38 .$. ,kBi7DdH:G<IIGIIGGEGJJIIJGGJHHEAJInJHCD  
chr11 26692847 C 38 .$. ,j@8fHDH<JJJJJJJDJJJIE@HICAIjoJJCE  
chr11 26692848 C 38 .....,^], 81dHDHDJJ@IJJHJJJEJiJHJJJJHGGIjqqJEEC  
chr11 26692849 A 39 .$. ,?1FBHFGJFJJJJHJJJJDDJJHIIHFDGJiQIHCEEC  
chr11 26692850 A 37 .....,jFDHFHH@IJFJICJJJJGIIICBCJJqJG7FEA  
chr11 26692851 A 37 .....,efBFCHH?IJJEIJCJiJJJGHID<=C=JiJqJG=FFD  
-UU-----F1 tumor.mpileup 1% L?? (Fundamental)-----
```

Pileup format

Describes base-pair information at each position

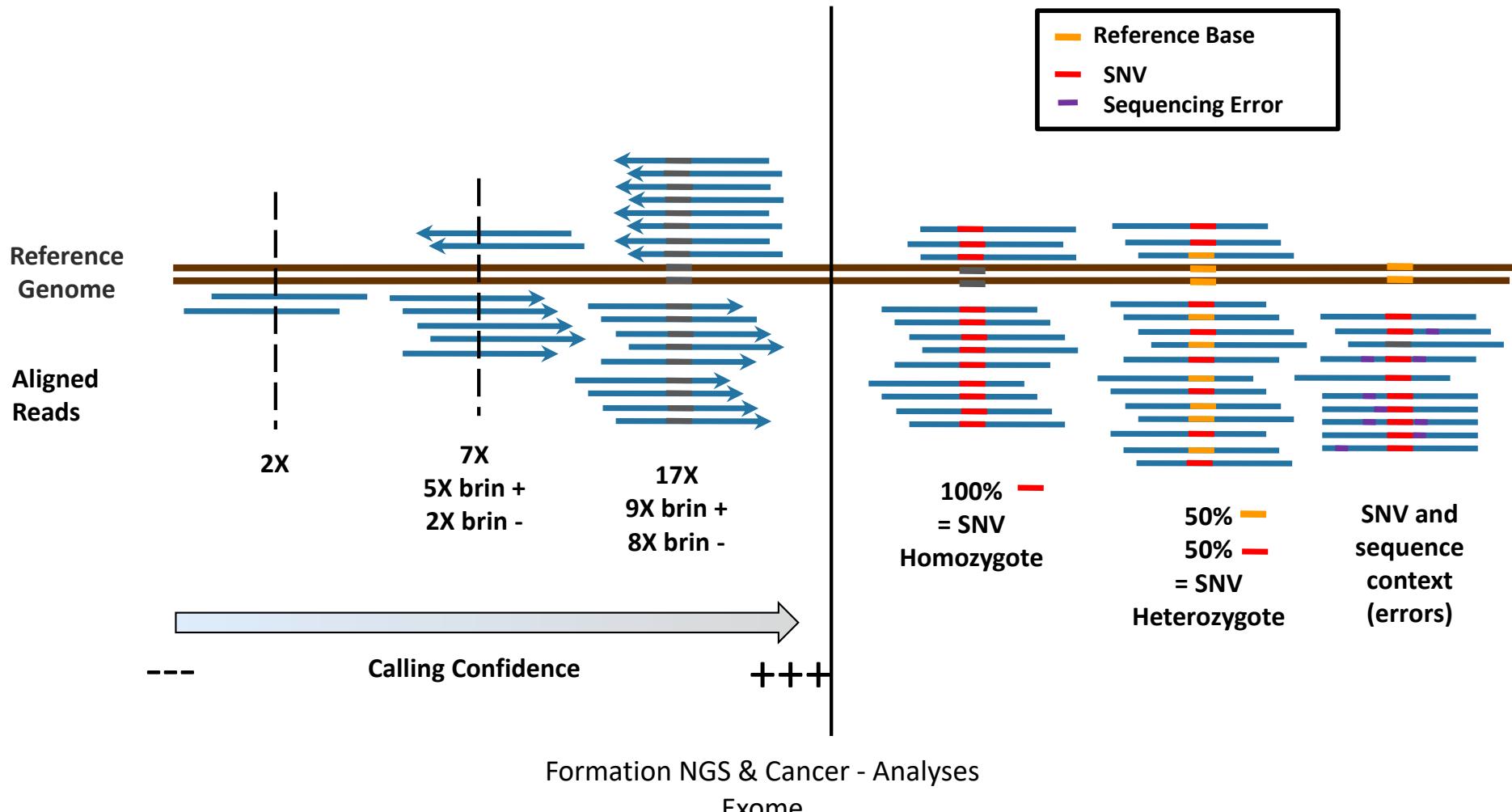
X2 for tumor & normal





Variant calling criteria

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



Factors to consider for Variant Calling

- 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
 - SNVs position within the 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
- Calling an indel:
 - Higher stringency (and Sanger validation often needed)

VarScan2

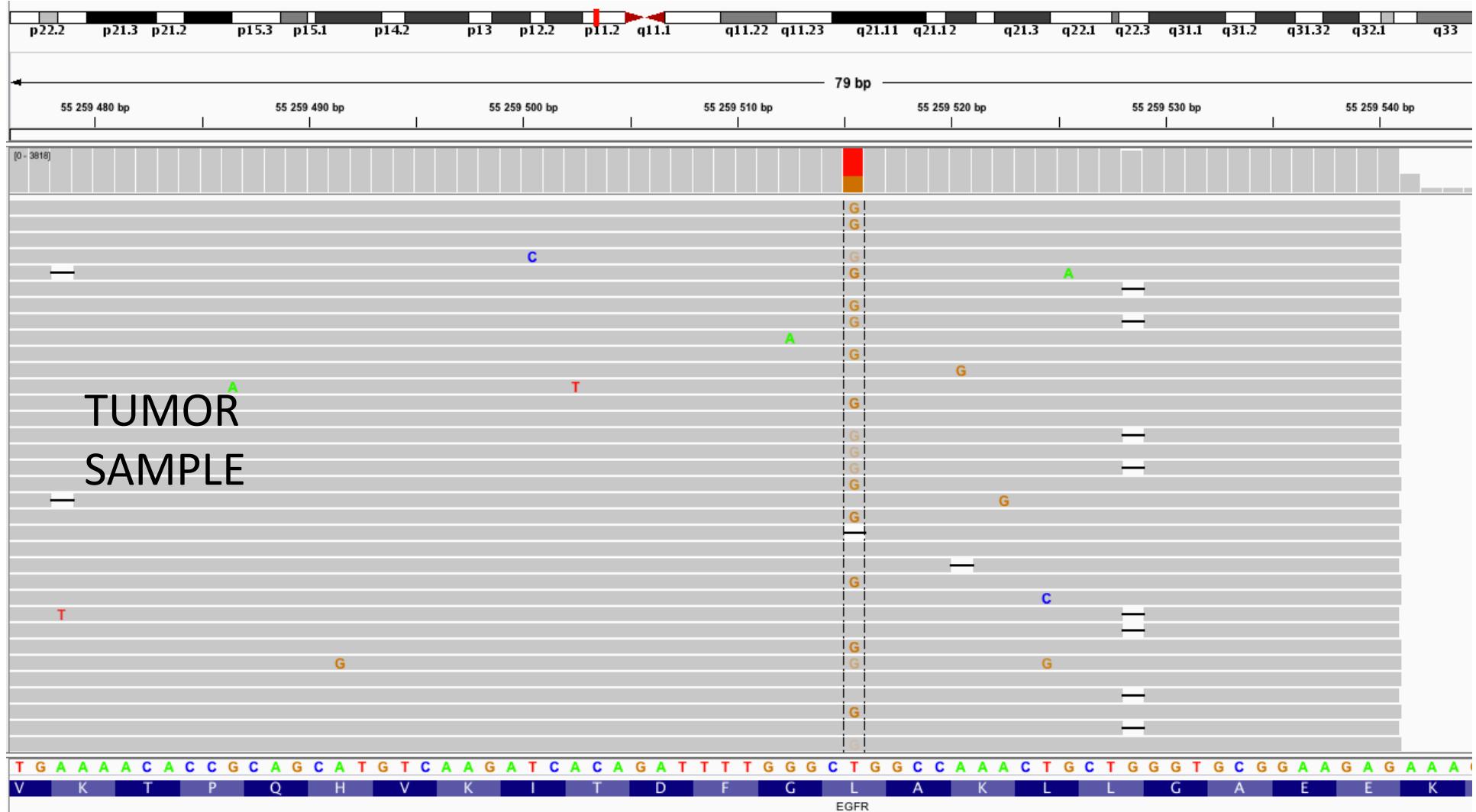
- Mutation caller written in **Java** (no installation required) working with **Pileup files** of Targeted, Exome, and Whole-Genome sequencing data (DNAseq or RNAseq)
- **Multi-platforms:** Illumina, SOLiD, Life/PGM, Roche/454
- Detection of different kinds of Germline SNVs/Indels (classical mode):
 - 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

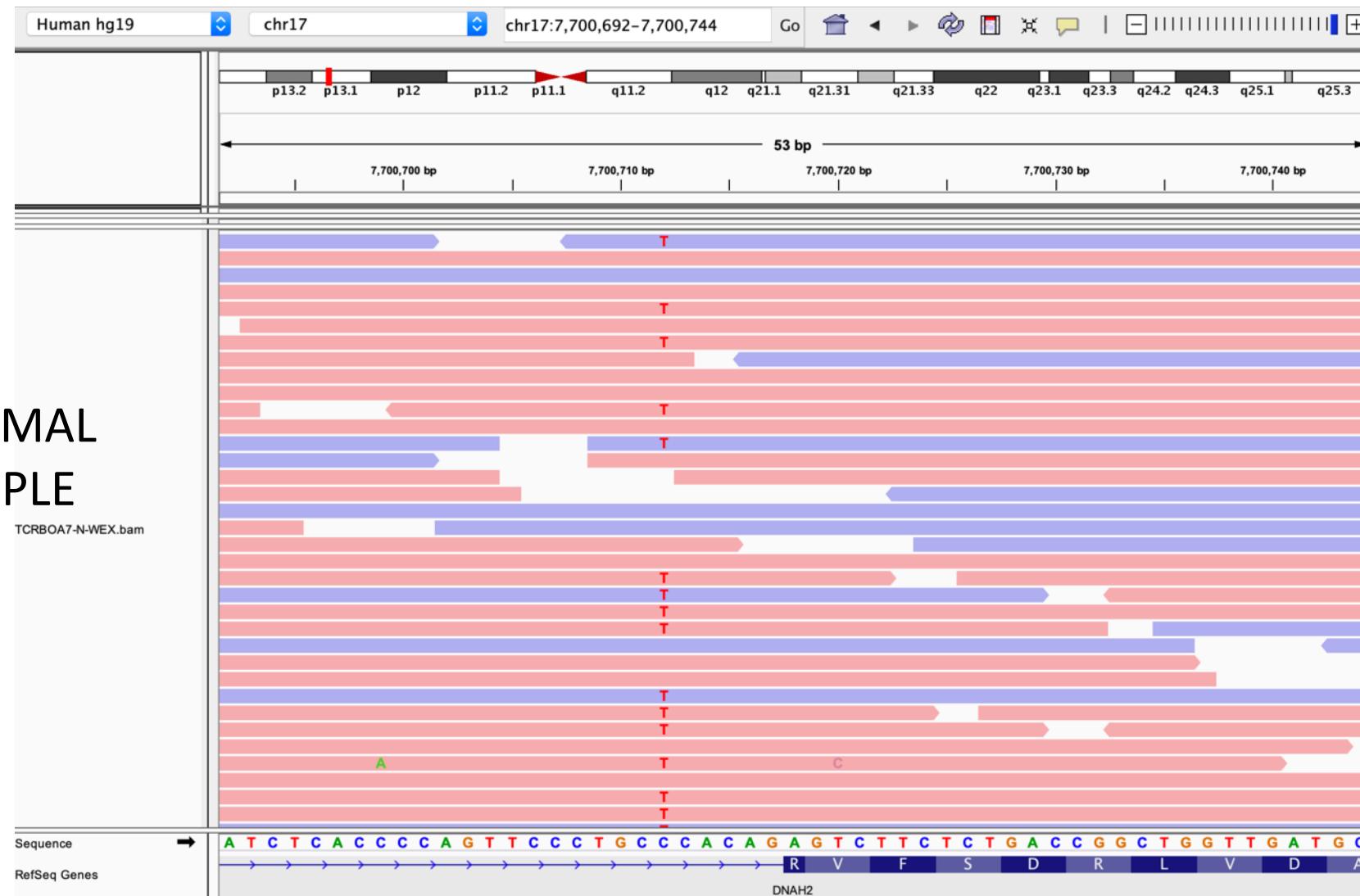
- VarScan uses a robust **heuristic/statistic** approach to call variants that meet desired thresholds for read depth, base quality, variant allele frequency, and statistical significance
- Stead *et al.* (2013) compared 3 different **somatic callers** : MuTect, Strelka, VarScan2
 - **VarScan2 performed 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**

Somatic calling

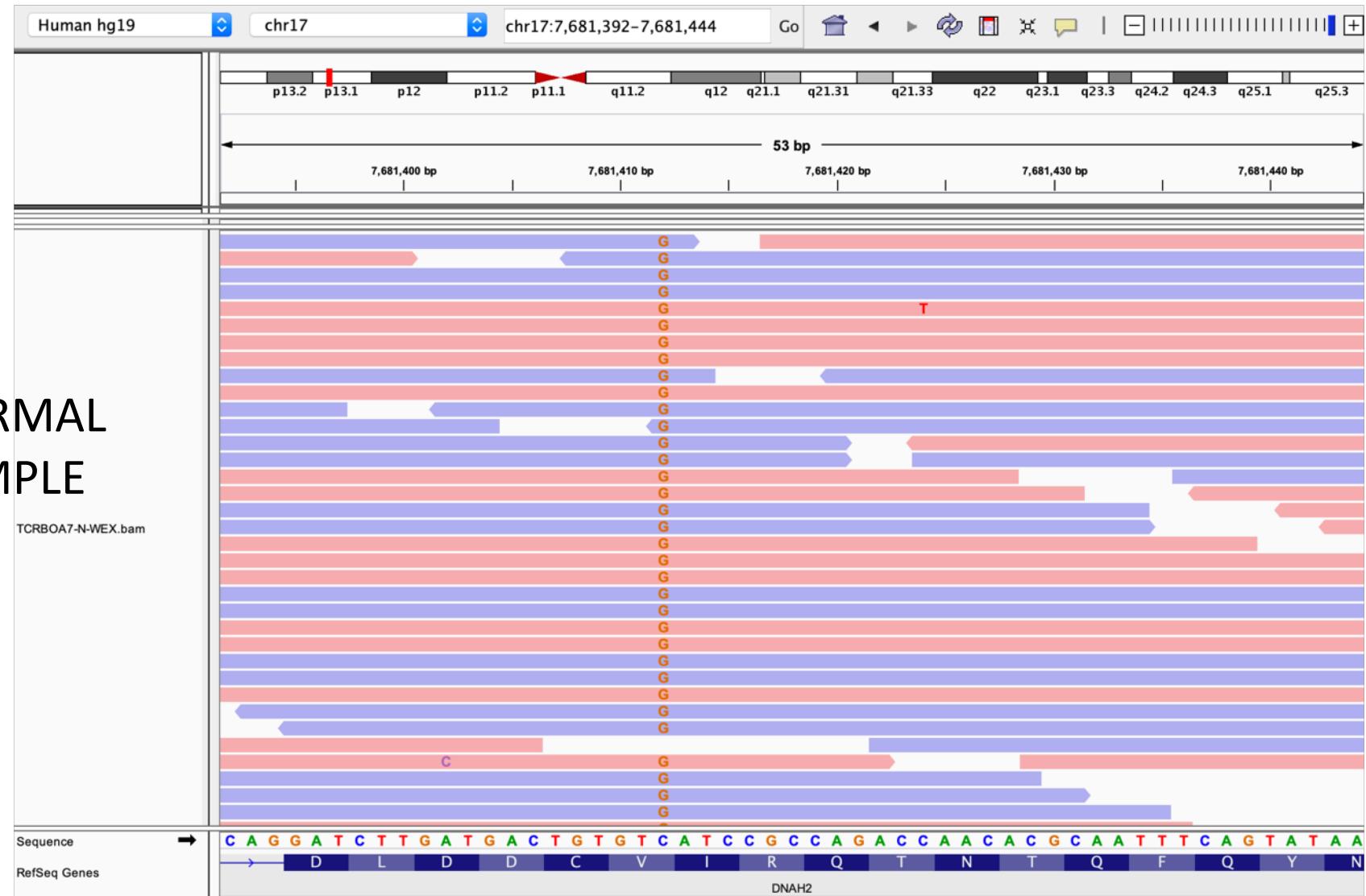
Un variant tumoral: polymorphisme ou mutation?



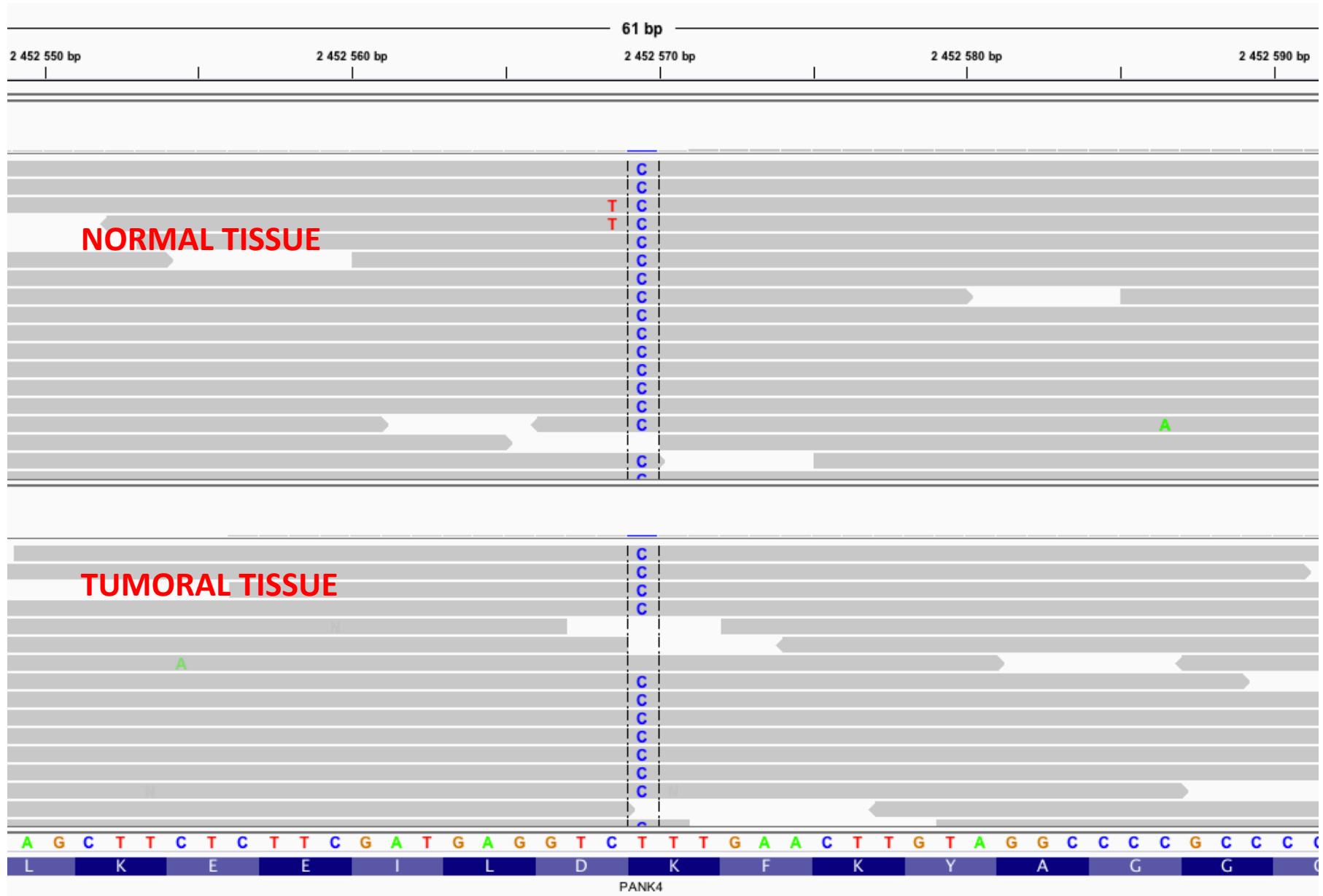
Un polymorphisme (SNP) hétérozygote



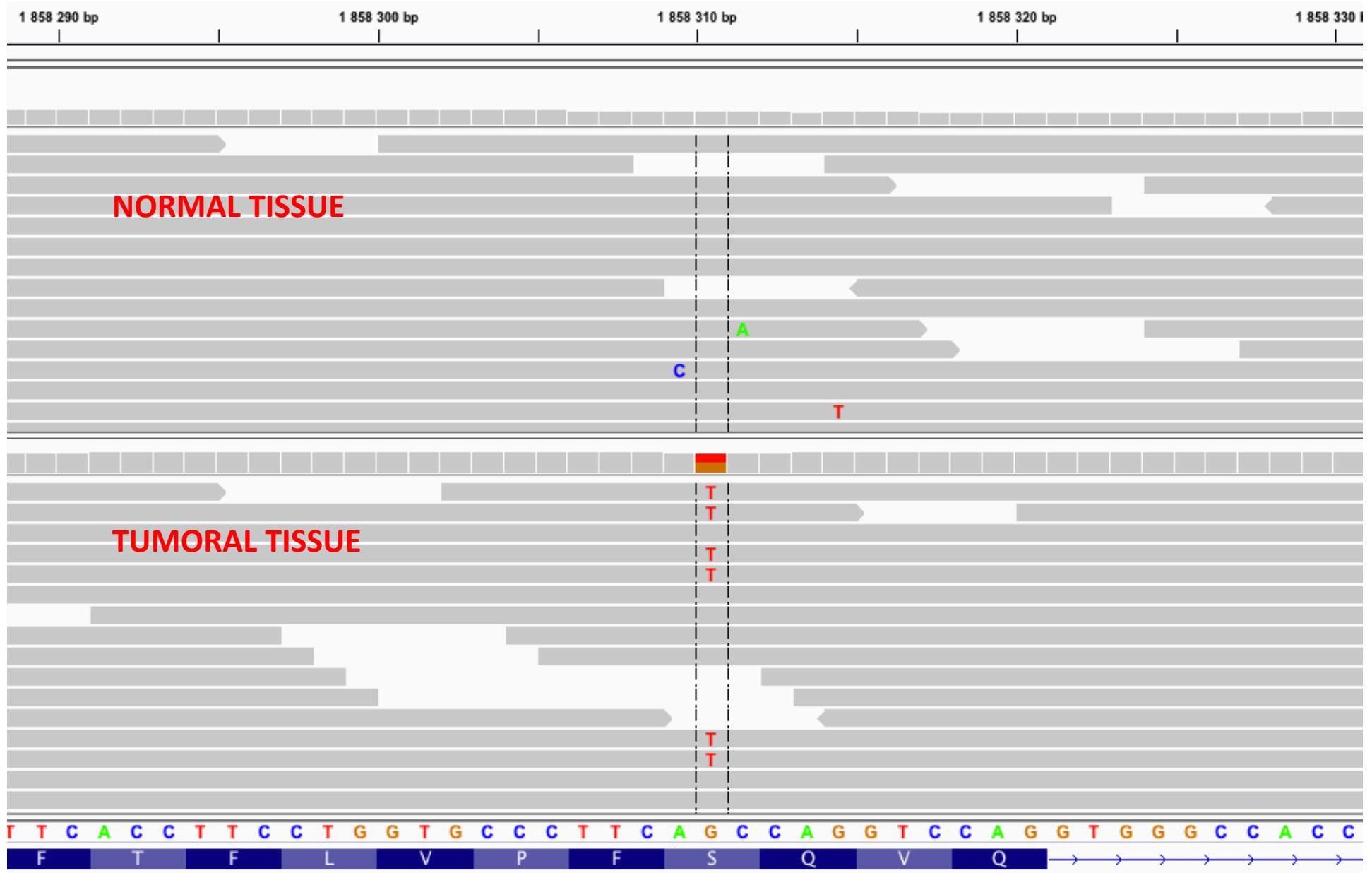
Un polymorphisme (SNP) homozygote



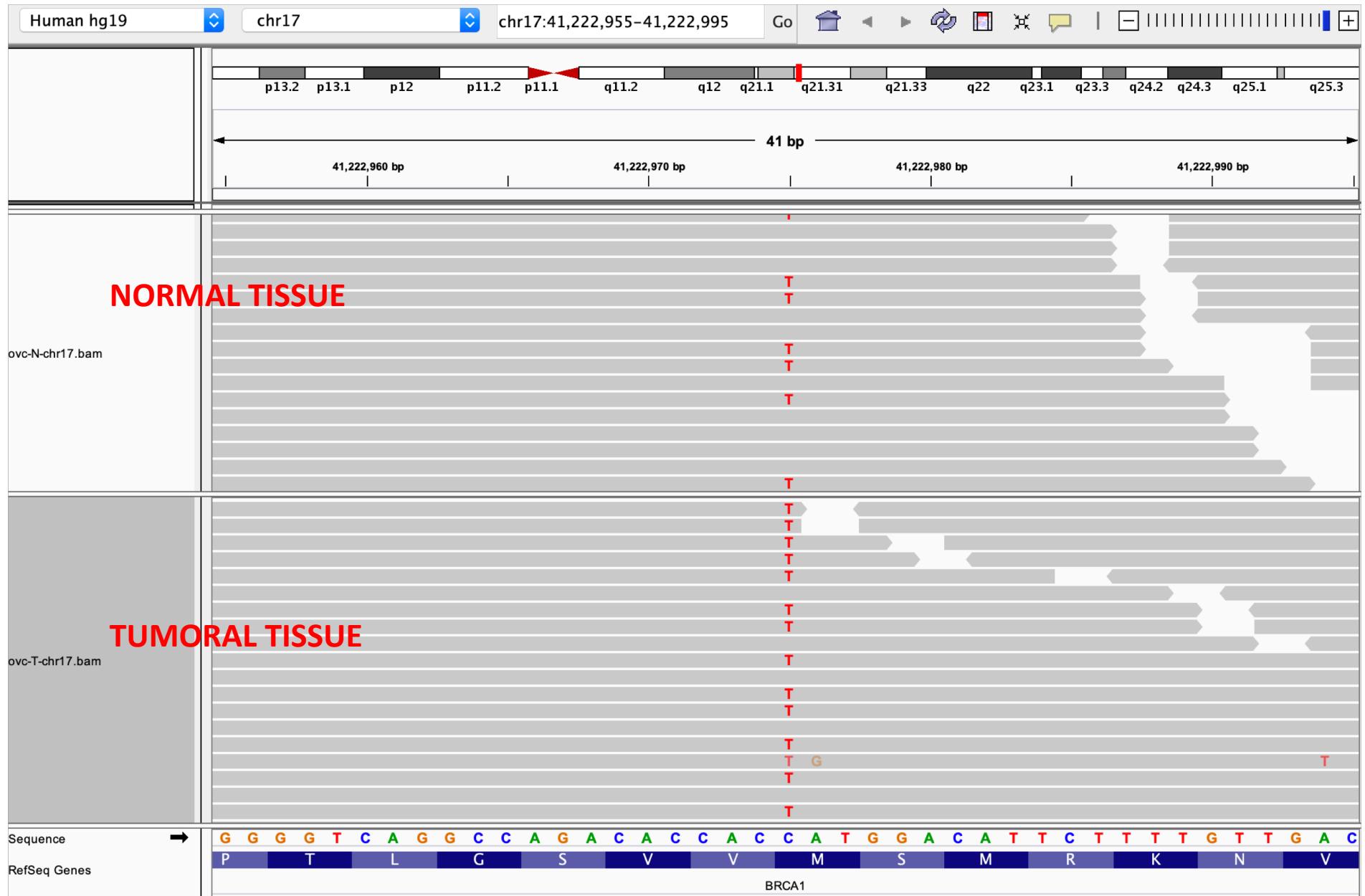
Un polymorphisme vu dans N et T



Une mutation somatique



Une LOH (loss of heterozygosity)



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:

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

The diagram illustrates the logic for determining if a variant is somatic or LOH based on allele counts in normal (N) and tumor (T) samples.

Alleles

	Ref	Var
N	8	0
T	6	7

Somatic Call: A red arrow points from the table to the text "Somatic".

	Ref	Var
N	4	4
T	8	1

LOH Call: A red arrow points from the table to the text "LOH".

Format VCF

```

##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">
#CHROM POS ID REF ALT QUAL FILTER INFO
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G

```

mandatory

Optional header: meta-data about available annotation

mandatory

samples

deletion

Insertion
(2 events here)

NS: number of samples with data
DP: combined depth
AF: allelic fraction
AA: ancestral allele

GT: genotype (0=ref, 1=alt)
GQ: genotype quality
DP: read depth

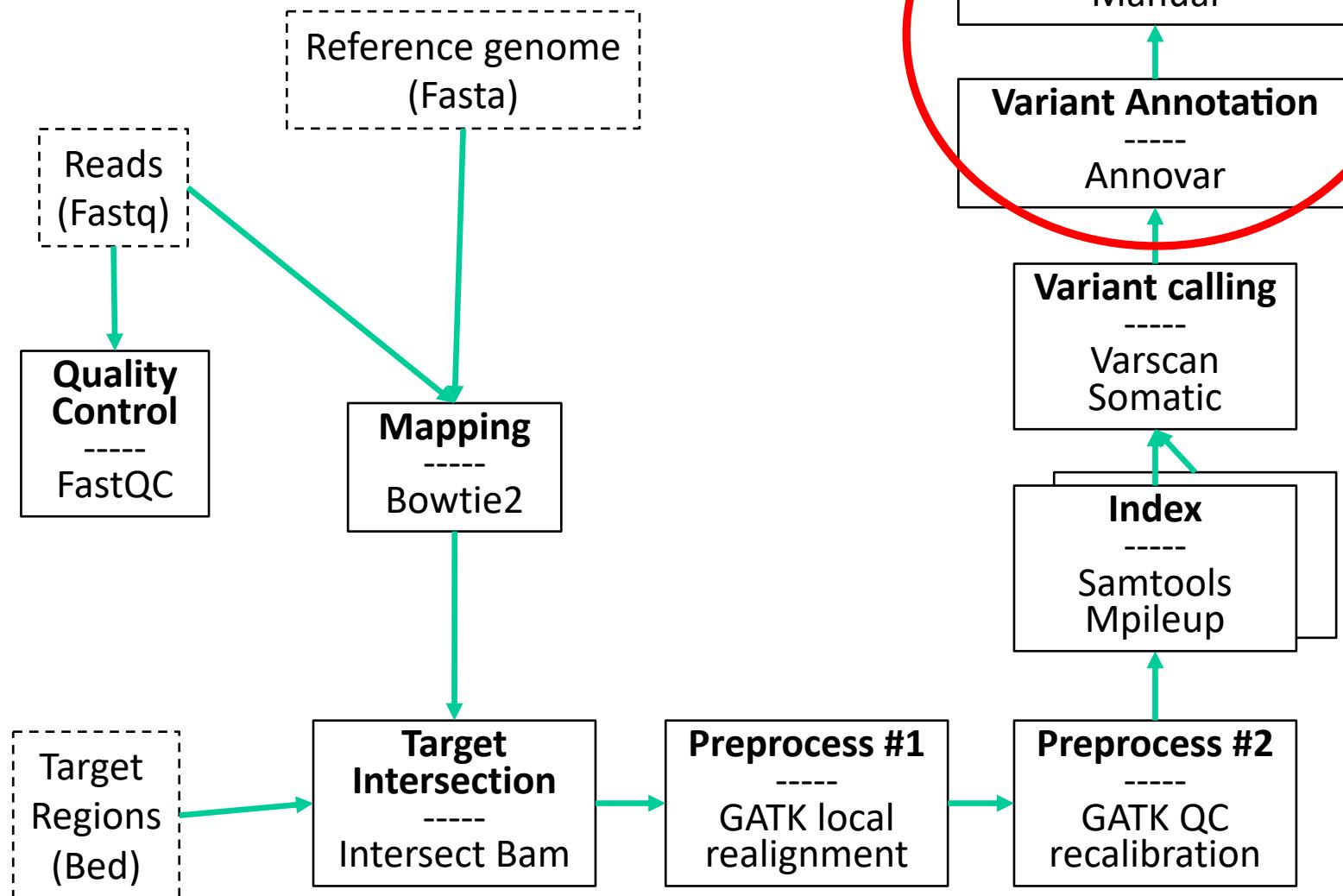
	FORMAT	NA00001	NA00002	NA00003
	GT:GQ:DP:HQ	0 0:48:1:51,51	1 0:48:8:51,51	1/1:43:5:...
	GT:GQ:DP:HQ	0 0:49:3:58,50	0 1:3:5:65,3	0/0:41:3
	GT:GQ:DP:HQ	1 2:21:6:23,27	2 1:2:0:18,2	2/2:35:4
	GT:GQ:DP:HQ	0 0:54:7:56,60	0 0:48:4:51,51	0/0:61:2
	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

Alternative: 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



Different types of SNVs

- SNVs and short indels are the most frequent events:
 - Intergenic
 - Intronic
 - *cis*-regulatory
 - splice sites
 - frameshift or not
 - synonymous or not
 - benign or damaging etc...
- Example of SNV one want to pinpoint:
 - 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

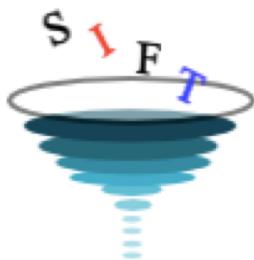
Takes Multi sample VCF (Tumor &+normal samples)

- 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 file

Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	ExonicFunc.refGene	AAChange.refGene	1000g2012apr_all	snp137	cosmic68	esp6500_all	LJB_PhylоП	LJB_PhylоП_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				

Formation NGS & Cancer - Analyses
Exome

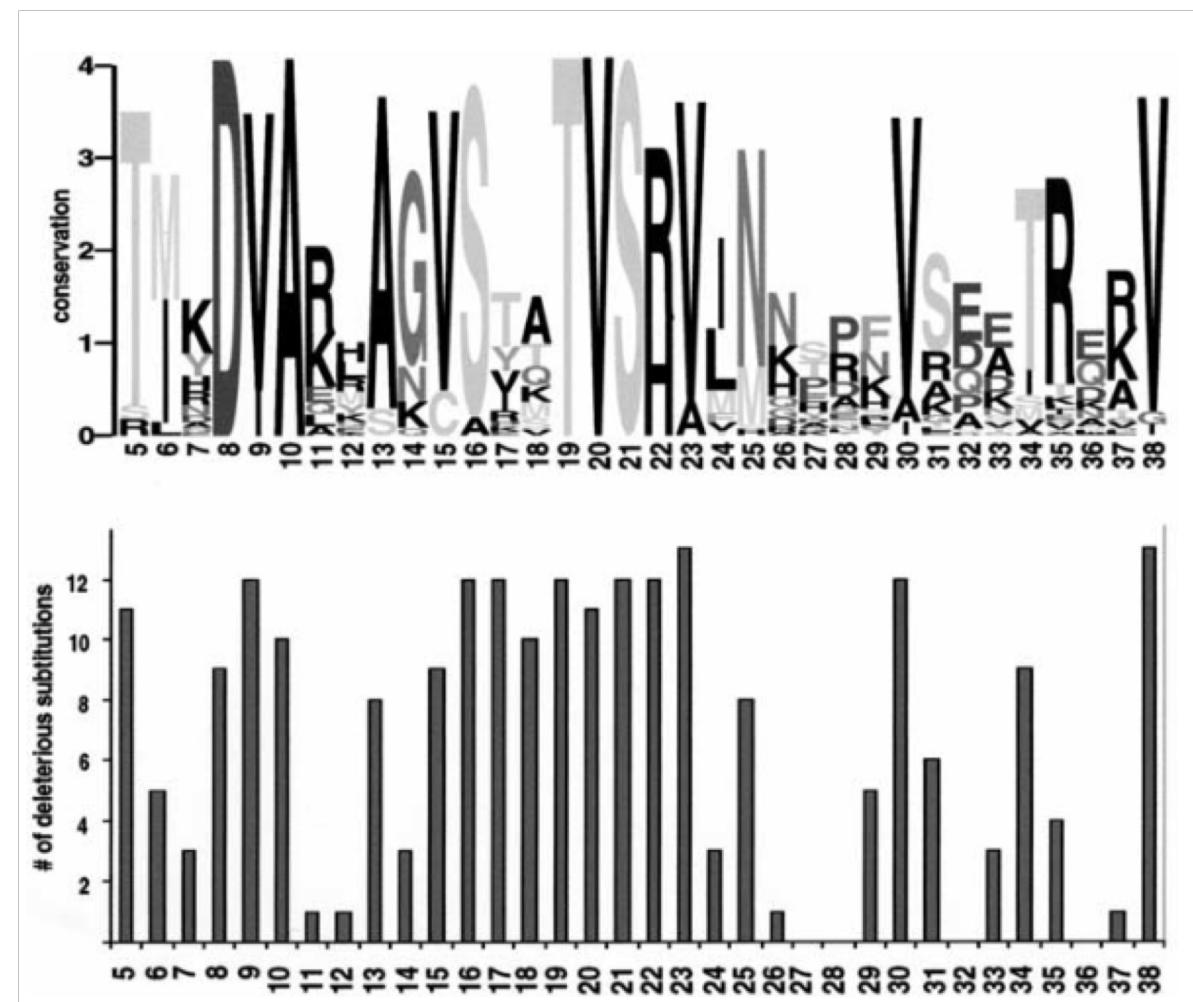


Sorting Intolerant From Tolerant

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)

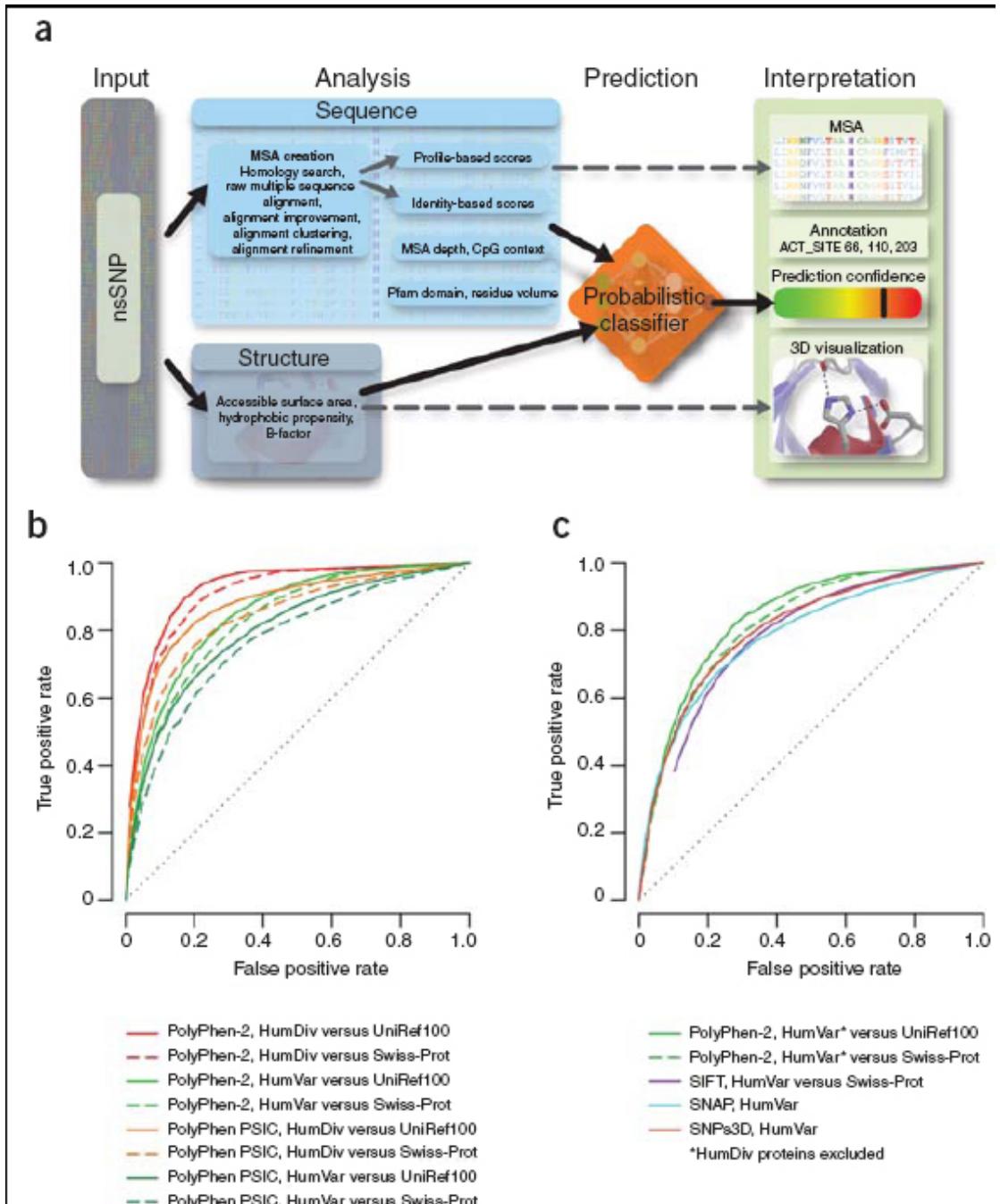


PolyPhen2

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.

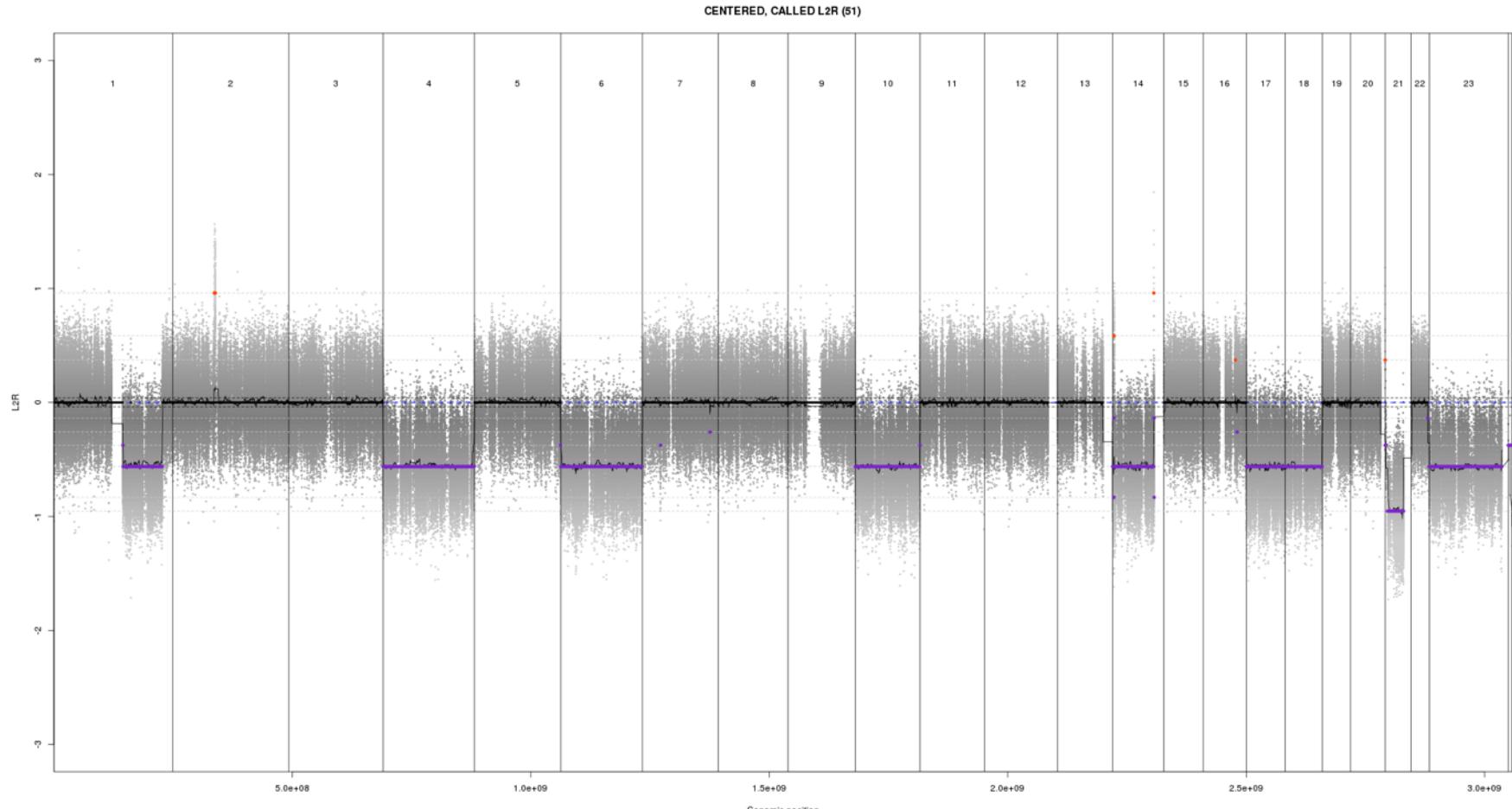
Classe en: **Benign**,
Possibly damaging, or
probably **Damaging**



Coverage & Allelic Frequencies For CNV detection

Detection of copy-number variations

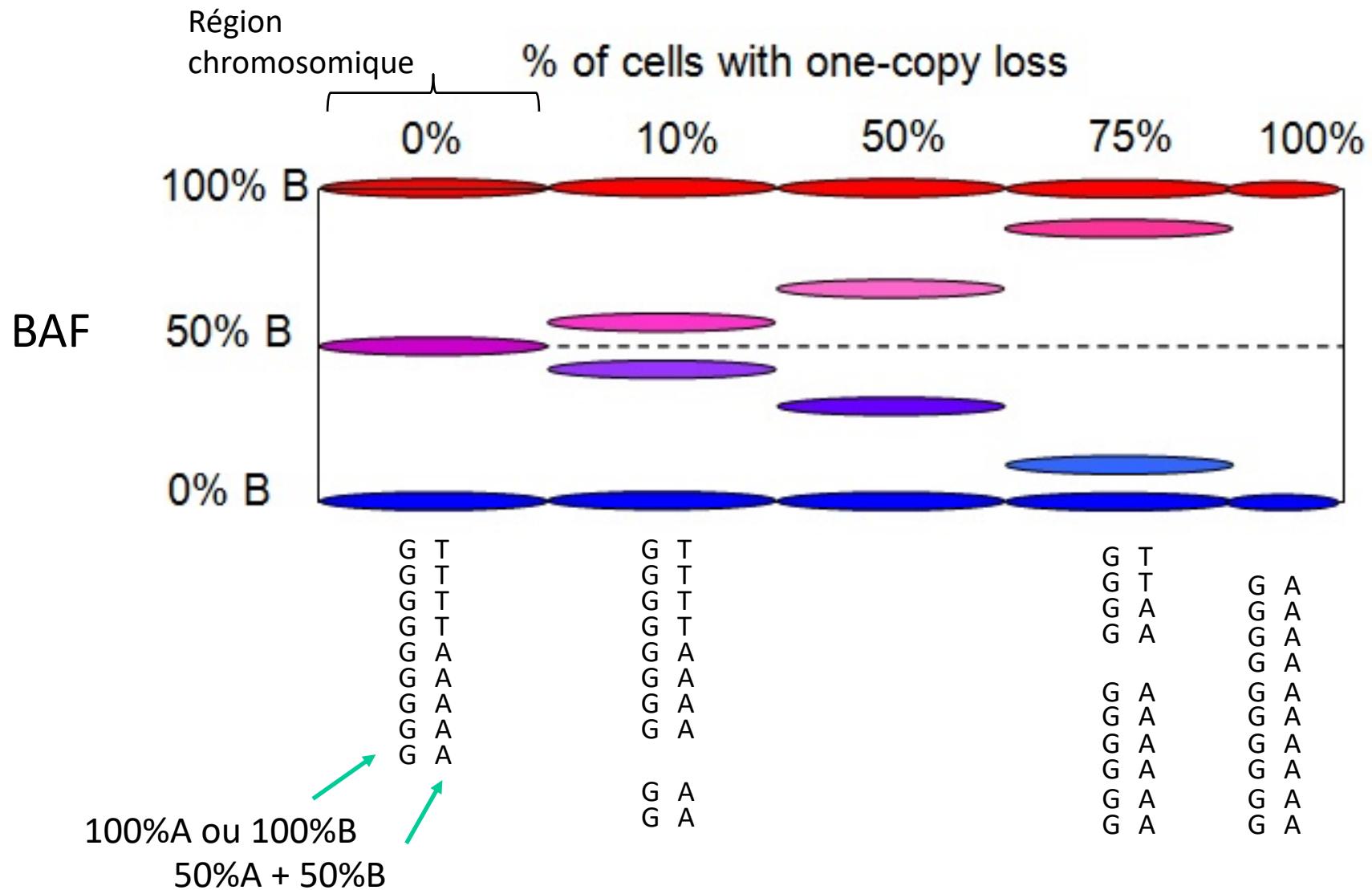
Are there any copy-number alteration (gain or loss of chromosomal regions, amplifications ...) that could explain tumorigenesis ?



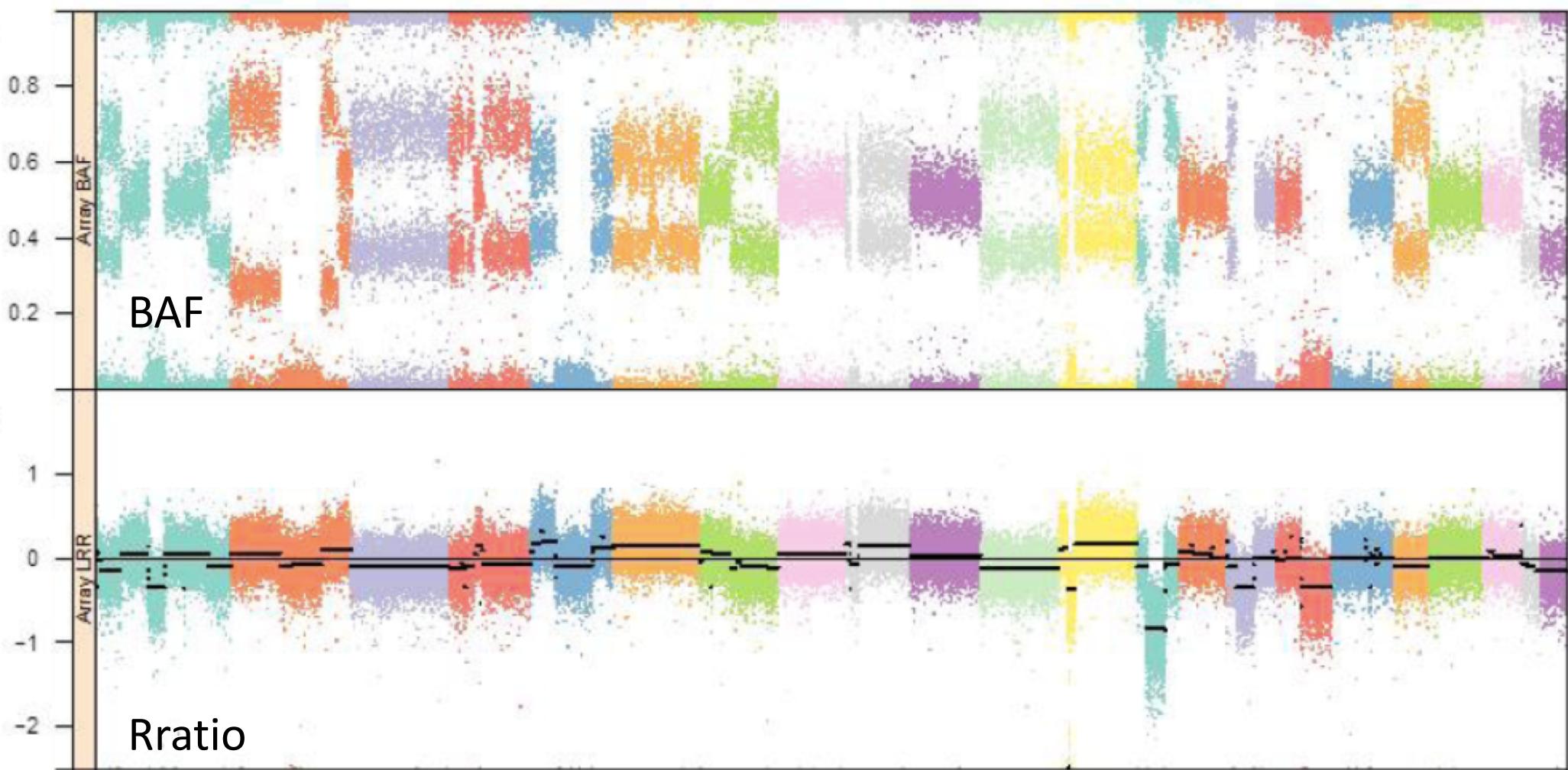
La fraction allélique

- 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

Cellularité et Fréquence Allelique



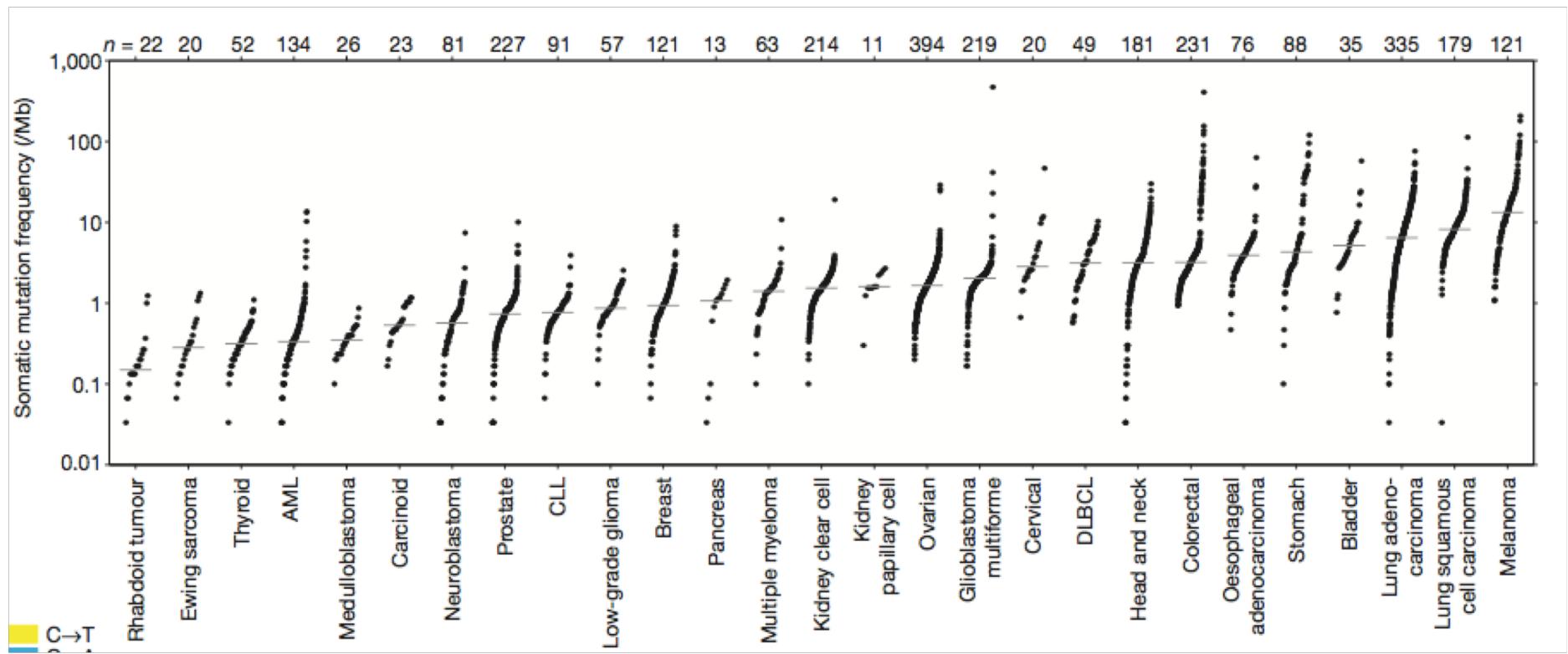
Segmentation et fréquence allélique



R ratio=utilisé en CGH, =couverture en NGS

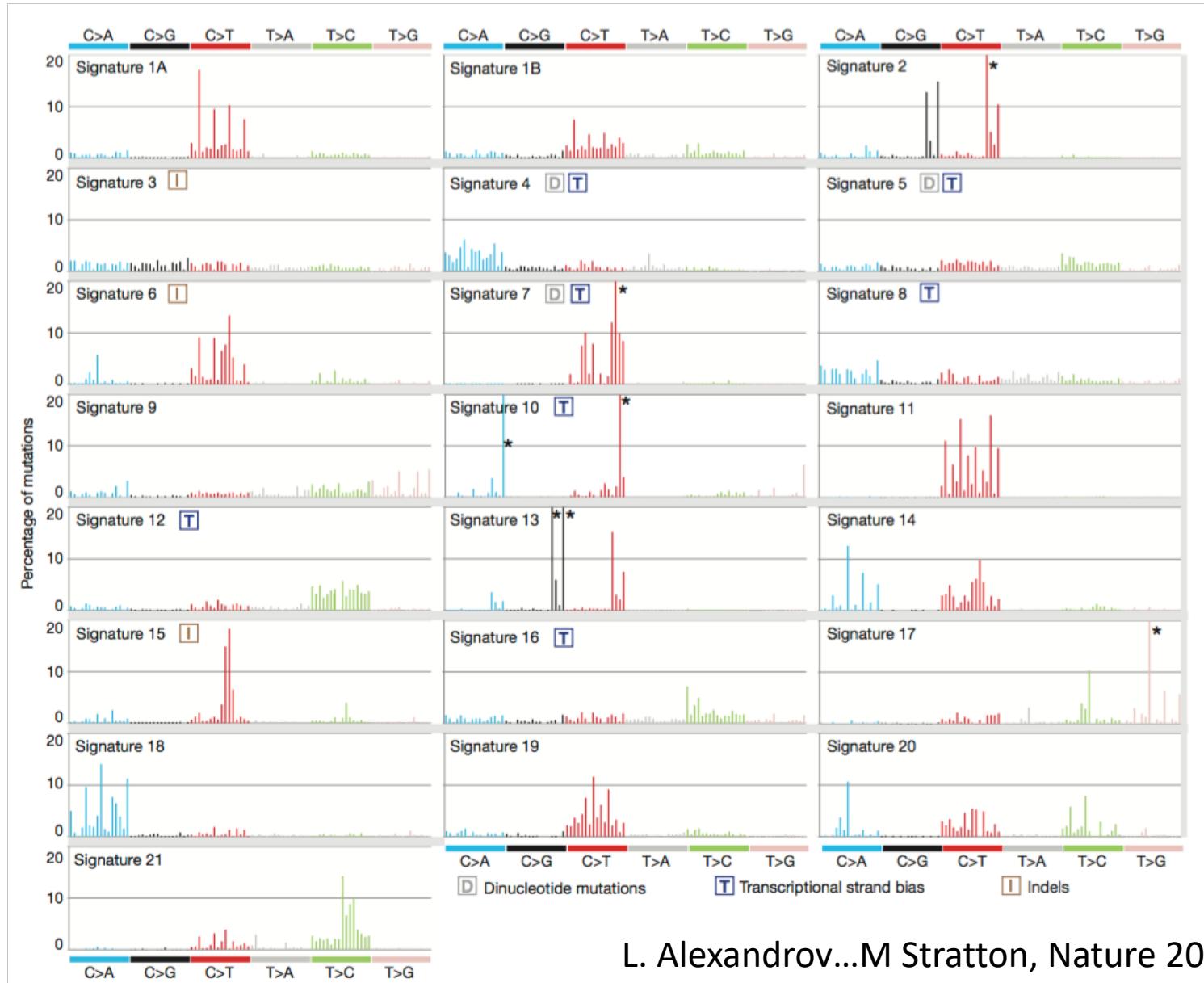
Scott et al. Gene 2014

2013: premières études pan-cancer WGS

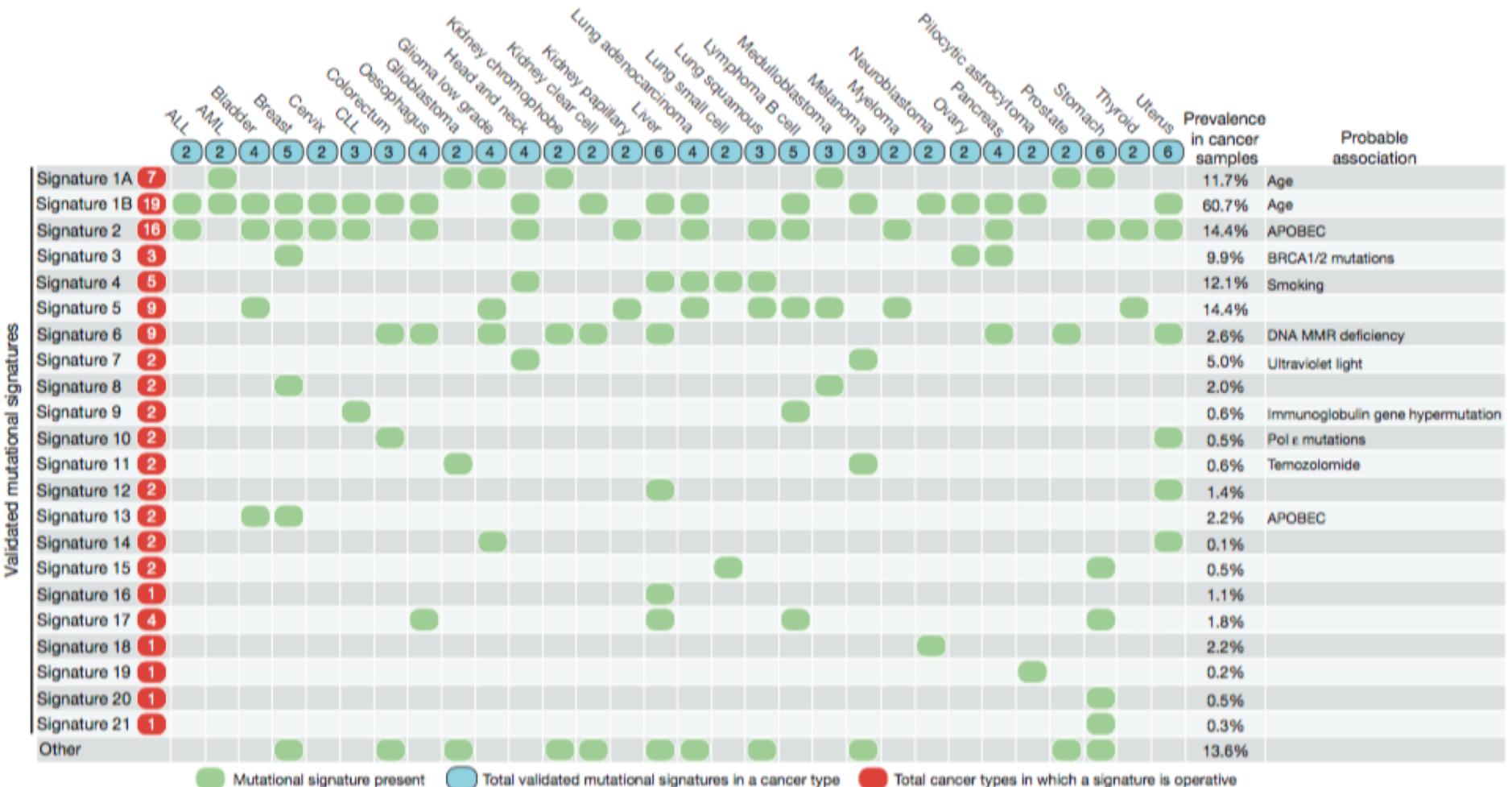


Alexandrov...M Stratton, Nature 2013.
Lawrence et al. Nature 2013.

Les Signatures Mutationnelles



Signatures et origine des tumeurs



Annexes

- Formats de fichiers NGS

Format fasta

*.fa , *.fasta

```
>identifiant1 commentaire libre
CAGCATCGATCGTCGGCGATGCATGCGGATGCTAGCTGATCACGATGC
CGCATGCTAGTCAGGCAGGGATATTATTAGCAGGTATCGGATGA
CAGCATTACGGCGGGAGTGCTATTATTATGAGCGGCGAT
>identifiant2 commentaire libre
CAGGCAGGTTCTTATTATATCGGCGGGCGGAGGCAGGCGATGCATC
CAGTGCAGTACGCTAGTCAGCGATGCATTATGACTGACTCAGTTT
CCCGCTAGCTATGCTATGCTATTGATCGATTGAGCTGATCTGGC
CAGCTATGCTTAGTA
```

Format fastq

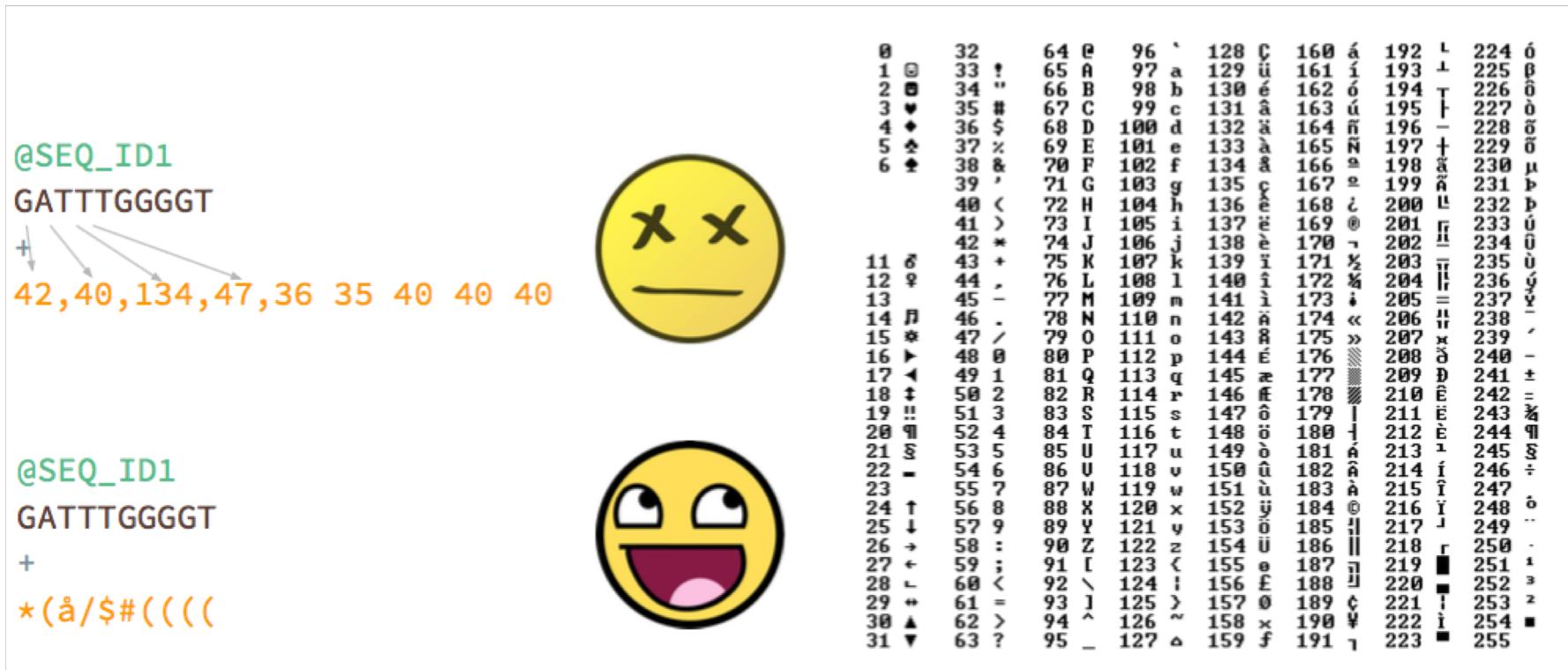
Descriptif du read (position sur la piste de séquençage, taille,...)

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Qualité (probabilité que la base soit correcte) encodé par code ASCII

Qualité dans le format fastq

$$\text{Qualité} = -10 \log_{10}(P_{\text{erreur}})$$



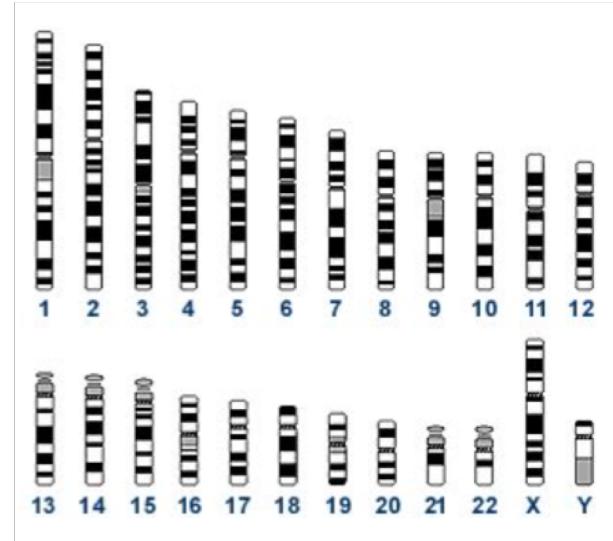
Fichiers de régions

Coordonnées génomiques indiquant une région du génome

<chromosome>:<start>-<end>
chr7:117465784-117715971

Formats de régions

- BED
- GTF/GFF: annotation de features dans le génome
- SAM/BAM: alignement de reads de séquence sur le génome
- VCF: variant calling file



Format bed

obligatoire		<i>name</i>	<i>score</i>	<i>strand</i>	<i>Thick start</i>	<i>Thick end</i>	<i>color</i>	
chr7	127471196	127472363	Pos1	0	+	127471196	127472363	255,0,0
chr7	127472363	127473530	Pos2	0	+	127472363	127473530	255,0,0
chr7	127473530	127474697	Pos3	0	+	127473530	127474697	0,255,0
chr7	127474697	127475864	Pos4	0	+	127474697	127475864	255,0,255

Attention

Le premier nucléotide est numéroté 0.

end - start = taille de la séquence



Format SAM/BAM

Rappel BAM:

@RG	ID:group1	SM:1425_CD34	PL:ILLUMINA	LB:lib1	PU:unit1				
@PG	ID:bwa	PN:bwa	VN:0.7.12-r1039	CL:bwa mem -M -t 2 -A 2 -E 1 -R	@RG\tID:group1\tSM:1425_CD34\tPL:ILLUMINA\tLB:lib1\tPU:unit1 /root/myd				
ERR166338.13782800	83	chr13	32890449	60	101M	=	32890343	-207	GGGACTGAATTAGAATTCAAACAAATTTCAGCGCTT
ERR166338.13782800	163	chr13	32890343	60	75M	=	32890449	207	CACTAGCCACGTTCGAGTGCTTAATGTGGCTAGTGGC
ERR166338.26716588	99	chr13	32890406	60	101M	=	32890553	222	AATGTTCCCACATCCTCACAGTAAGCTGTTACCGTTCCAG
ERR166338.26716588	147	chr13	32890553	60	75M	=	32890406	-222	TTGCAGACTTACCAAGCATTGGAGGAATATCGTAGGTAA
ERR166338.27259961	99	chr13	32890496	60	101M	=	32890558	137	ACCTCAGTCACATAATAAGGAATGCATCCCTGTGTAAG
ERR166338.27259961	147	chr13	32890558	60	75M	=	32890496	-137	GACTTATTTACCAAGCATTGGAGGAATATCGTAGGTAA
ERR166338.63037998	99	chr13	32890496	60	101M	=	32890558	137	ACCTCAGTCACATAATAAGGAATGCATCCCTGTGTAAG
ERR166338.63037998	147	chr13	32890558	60	75M	=	32890496	-137	GACTTATTTACCAAGCATTGGAGGAATATCGTAGGTAA

Diagram illustrating the fields in a BAM header line:

- read ID: Points to the first field (ID:group1).
- flag: Points to the second field (0).
- position: Points to the third field (chr13).
- CIGAR: Points to the fourth field (32890449).
- mapping qual.: Points to the fifth field (60).
- mate info: Points to the last field (GACTTATTTACCAAGCATTGGAGGAATATCGTAGGTAA).

Le champ CIGAR

Example:

52M36890N45M3S

REF : chr20



All Cigar operations

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

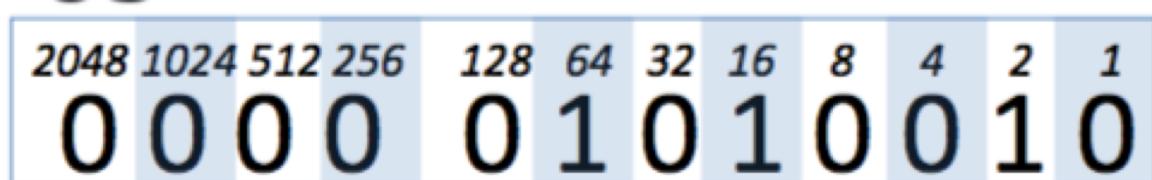
Les Flags SAM

Example:

- Decimal Flag Value

83

- Binary Flag Value



- To each bit corresponds a meaning

Bit	Description
1	0x1 template having multiple segments in sequencing
2	0x2 each segment properly aligned according to the aligner
4	0x4 segment unmapped
8	0x8 next segment in the template unmapped
16	0x10 SEQ being reverse complemented
32	0x20 SEQ of the next segment in the template being reverse complemented
64	0x40 the first segment in the template
128	0x80 the last segment in the template
256	0x100 secondary alignment
512	0x200 not passing filters, such as platform/vendor quality controls
1024	0x400 PCR or optical duplicate
2048	0x800 supplementary alignment

Format VCF

```

##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">
#CHROM POS ID REF ALT QUAL FILTER INFO
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G

```

mandatory

Optional header: meta-data about available annotation

mandatory

samples

deletion

Insertion
(2 events here)

NS: number of samples with data
DP: combined depth
AF: allelic fraction
AA: ancestral allele

	FORMAT	NA00001	NA00002	NA00003
GT:GQ:DP:HQ	0 0:48:1:51,51	1 0:48:8:51,51	1/1:43:5:...	
GT:GQ:DP:HQ	0 0:49:3:58,50	0 1:3:5:65,3	0/0:41:3	
GT:GQ:DP:HQ	1 2:21:6:23,27	2 1:2:0:18,2	2/2:35:4	
GT:GQ:DP:HQ	0 0:54:7:56,60	0 0:48:4:51,51	0/0:61:2	
GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3	

GT: genotype (0=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
20 - Tumor_Validation_Allele1	Secondary data from orthogonal technology. Tumor genotyping (validation) for 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
21 - Tumor_Validation_Allele2	Secondary data from orthogonal technology. Tumor genotyping (validation) for allele 2
22 - Match_Norm_Validation_Allele1	Secondary data from orthogonal technology. Matched normal genotyping (validation) for 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)