CRCM
Centre de Recherche
en Cancérologie de Marseille

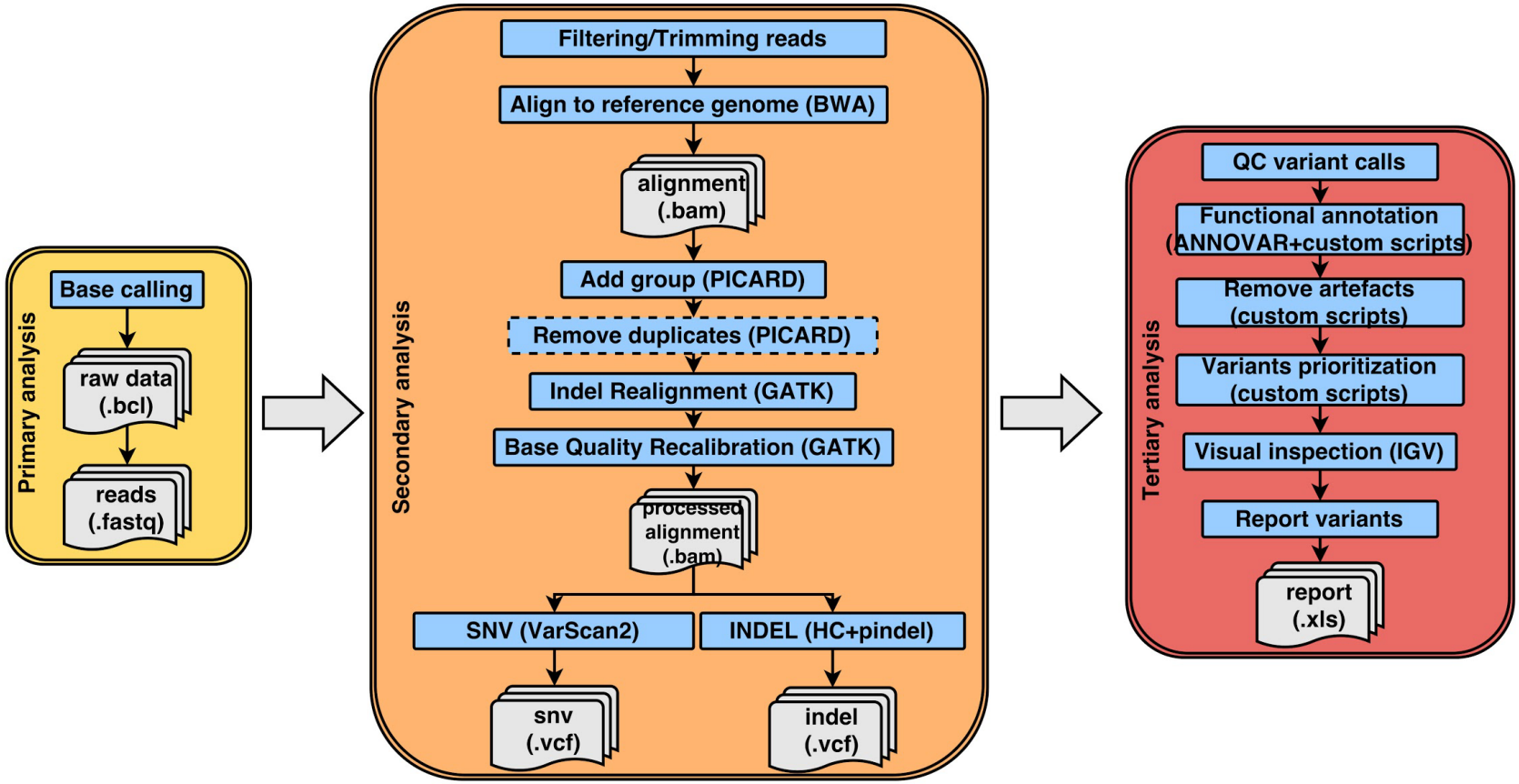
Les mardis de la technologie gourmande de DISC

Analyse bio-informatique appliquée à la détection de variants (NGS) Besoins et Solutions

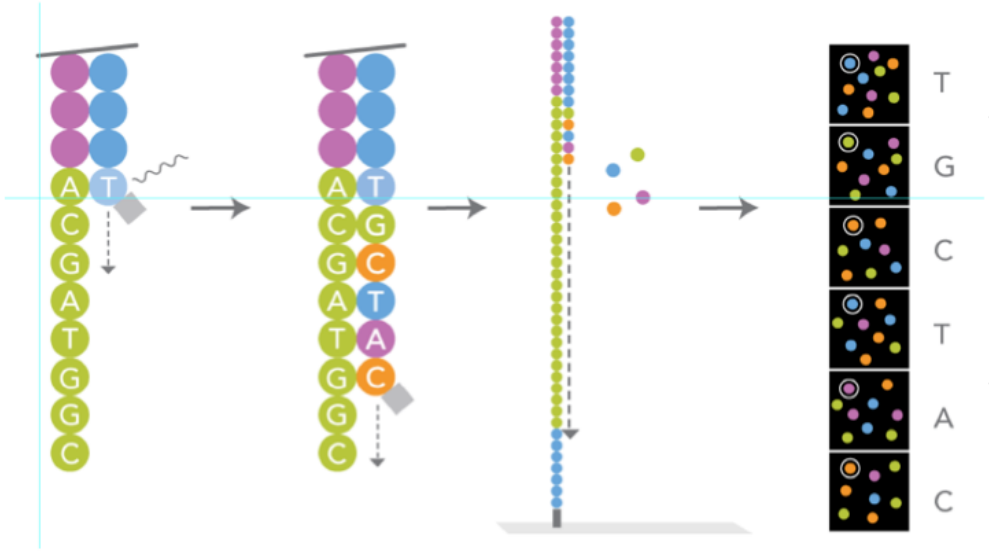
Arnaud Guille

Instituts thématiques Inserm Institut national de la santé et de la recherche médicale
INSTITUT PAOLI-CALMETTES
cnrs
Aix*Marseille université

Workflow


Primary analysis
Base calling
raw data (.bcl)
reads (.fastq)
Secondary analysis
Filtering/Trimming reads
Align to reference genome (BWA)
alignment (.bam)
Add group (PICARD)
Remove duplicates (PICARD)
Indel Realignment (GATK)
Base Quality Recalibration (GATK)
processed alignment (.bam)
SNV (VarScan2)
INDEL (HC+pindel)
snv (.vcf)
indel (.vcf)
Tertiary analysis
QC variant calls
Functional annotation (ANNOVAR+custom scripts)
Remove artefacts (custom scripts)
Variants prioritization (custom scripts)
Visual inspection (IGV)
Report variants
report (.xls)

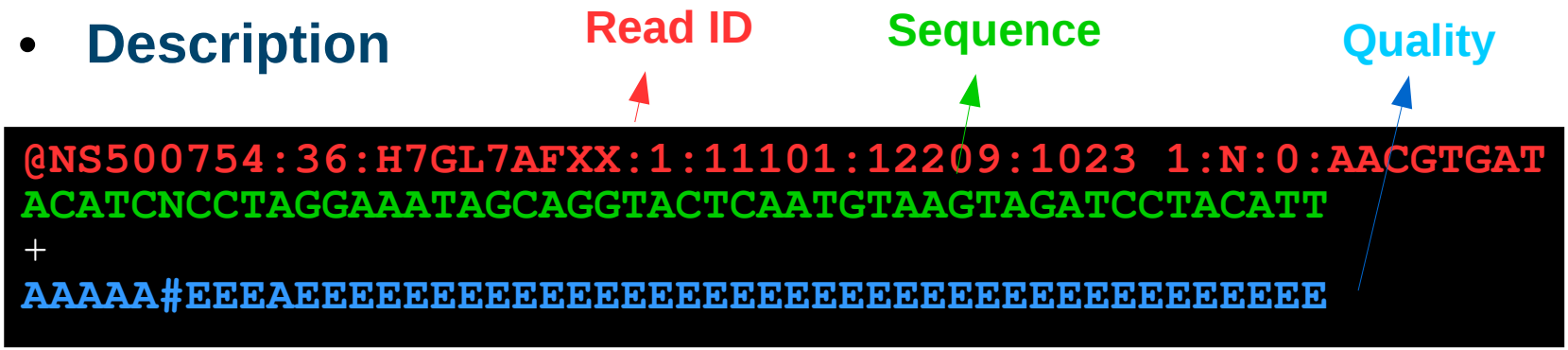
Analyse primaire : Base-calling



- Déconvolution du signal en séquence ATCG
- Assignation d'un score de qualité à chaque base (score phred)
- Généralement réalisé par le séquenceur lui même
- Fichier Fastq en sortie

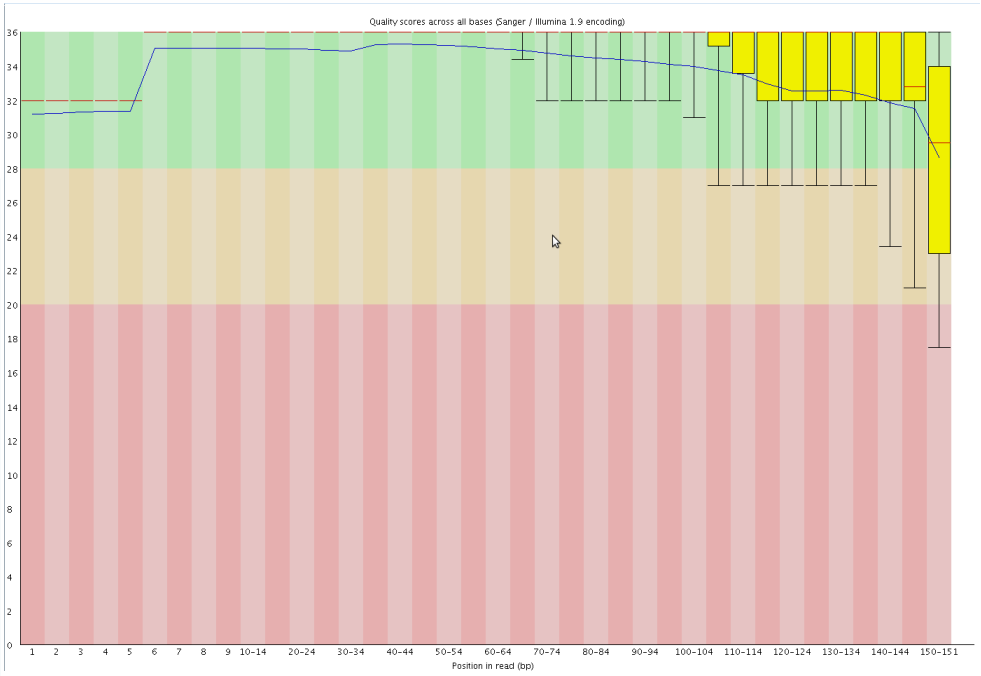
Format : Fastq

- **Description**



  - Inventé dans les années 90 par Jim Mullikin
  - Séquence et score codées avec un seul caractères ASCII
- **Score Phred**
  - Q = -10 * log10p
  - « A » <=> 65 <=> Q32 <=> p = 6,3e-4 (phred +33)

Analyse secondaire : Contrôle qualité et trimming

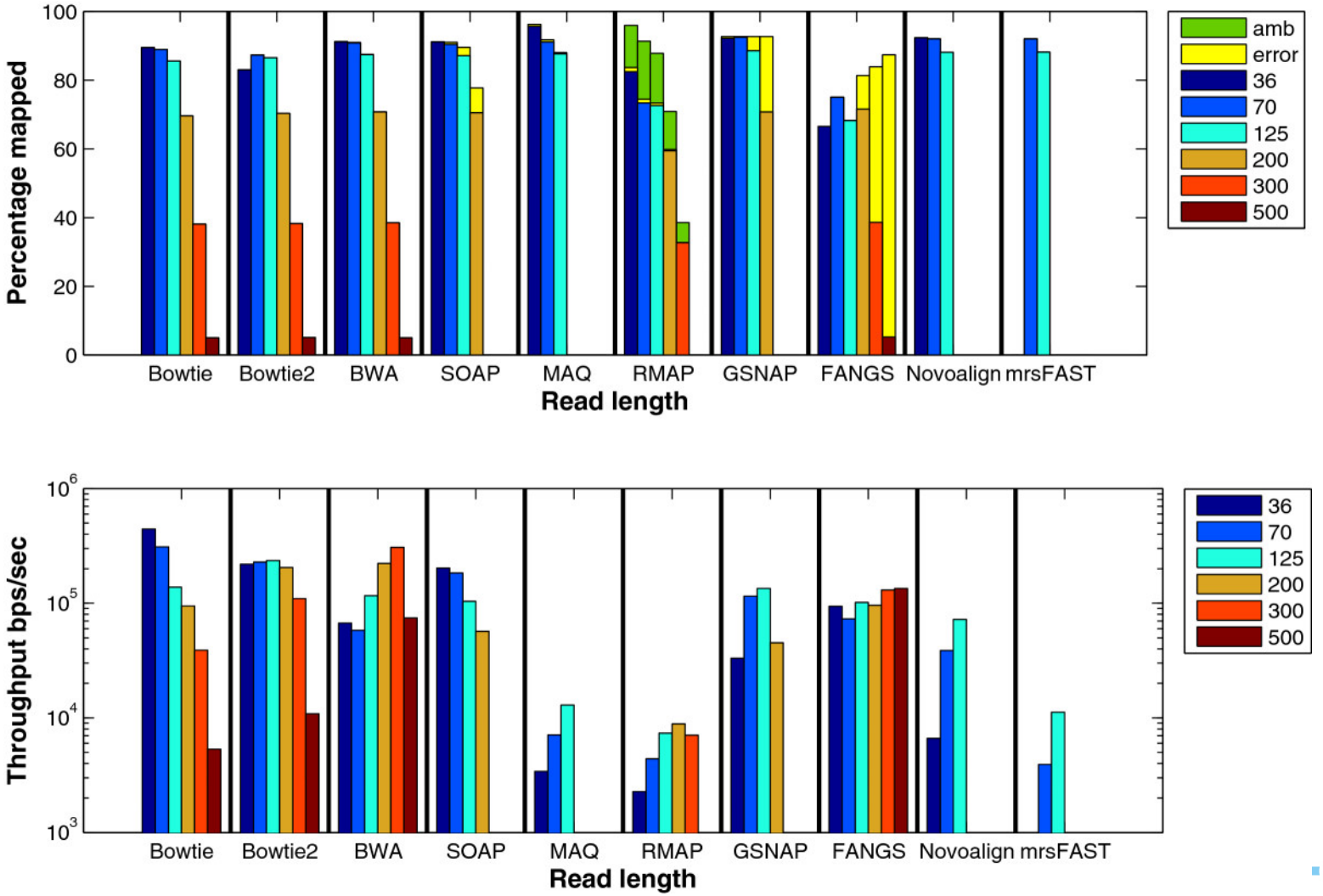


- Contrôle qualité des reads avec FastQC
- Les extrémités des reads sont souvent de moins bonne qualité
- Trimming des reads avec Trimmomatic

Analyse secondaire : L'alignement

	Bowtie	BWA	SOAP2	MAQ	RMAP	GSNAP	FANGS	Novoalign	mrFAST
Seed mm.	Up to 3	Any	Up to 2	Any	Any				
Non-seed mm.	QS	Count	Count	QS	Count	Count	Count	QS	Count
Var. seed len.	> 5	Any	> 28						
Mapping qual.		Yes		Yes				Yes	
Gapped align.		Yes	PE	PE		Yes	Yes	Yes	Yes
Colorspace	Yes	Yes		Yes				Yes	
Splicing						Yes			
SNP tolerance						Yes			
Bisulphite reads					Yes	Yes		Yes	Yes

Analyse secondaire : L'alignement



Hatem, A., Bozdag, D., Toland, A. E. & Catalyürek, Ü. V. Benchmarking short sequence mapping tools. BMC Bioinformatics 14, 184 (2013).

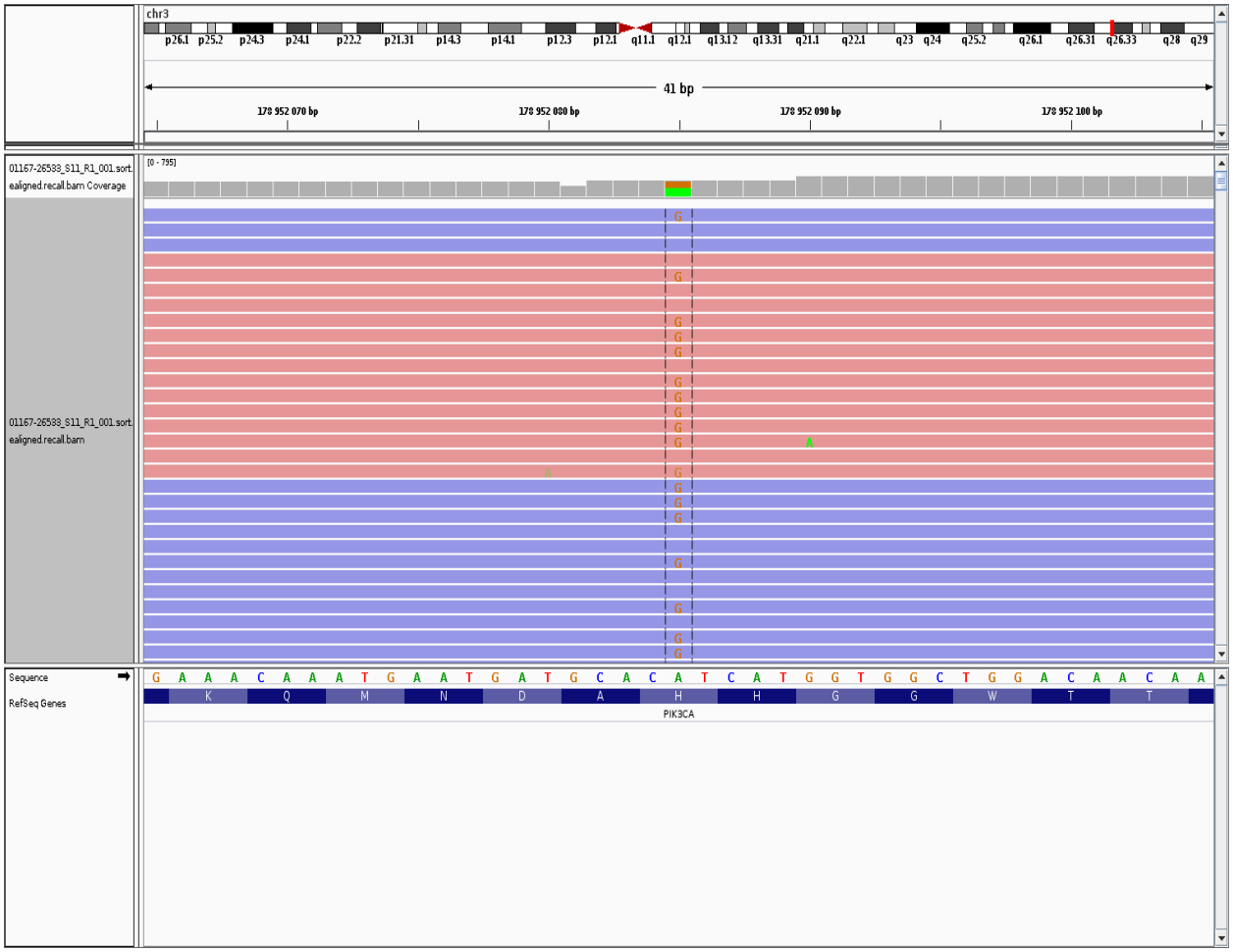
Format : SAM/BAM

- **SAM (Sequence Alignement Map)**

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

- **BAM**
  - Format binaire
  - Compact (~ 80 % plus petit)
  - Peut être indexé

Visualisation : SAM/BAM

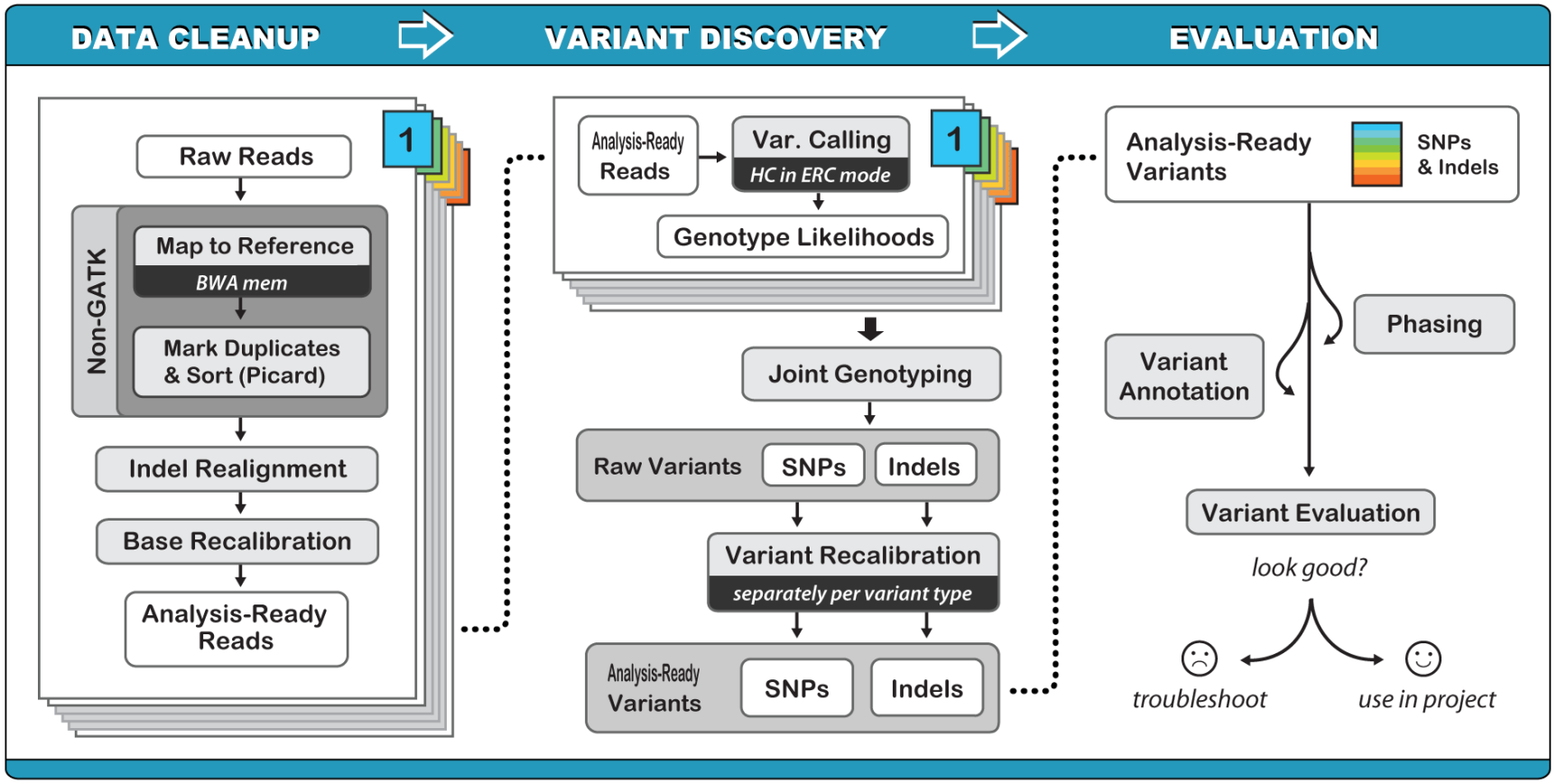


API/Outils de manipulation : SAM/BAM

- Samtools (C)
  - index, stats, view, sort, merge, mpileup, etc ….
- Picard tools (java)
  - AddOrReplaceReadGroups, SortSam, MarkDuplicates, etc ...
- Pysam (Python)

```
import pysam
samfile = pysam.AlignmentFile("ex1.bam", "rb")
pairedreads = pysam.AlignmentFile("allpaired.bam", "wb",
template=samfile)
for read in samfile.fetch():
     if read.is_paired:
               pairedreads.write(read)
pairedreads.close()
samfile.close()
```

Analyse secondaire : GATK Best Practices



Analyse secondaire : Réalignement autour des indels

- **Ce que dit le GATK :**

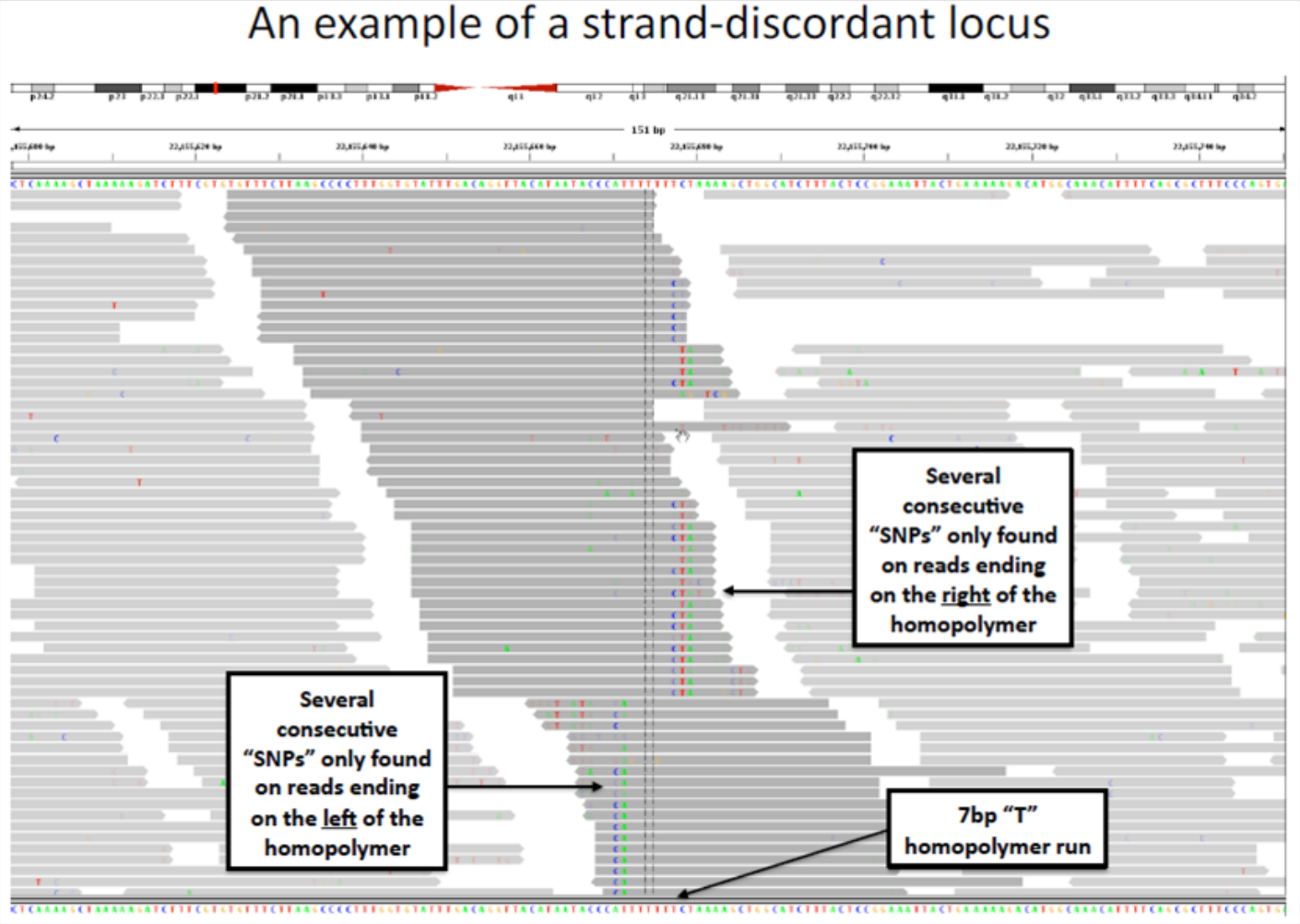
Answers

Geraldine_VdAuwera Posts: **9,634** Administrator, Dev admin
September 2013

I'm not familiar enough with BWA mem to give you a definitive answer, but basically the key point is that traditionally mappers like BWA only align one read at a time, which is why realignment around indels is necessary to provide consistent mapping -- the indel realigner sees all the reads in a region, not just one. So it depends if BWA mem has that added capability. I'm not aware that it does, so I would still recommend doing indel realignment according to our best practices, but I could be wrong.

Geraldine Van der Auwera, PhD

Analyse secondaire : Réalignement autour des indels



Analyse secondaire : Réalignement autour des indels

Analyse secondaire : Réalignement autour des indels

- **Comment ca marche :**
  - **Les régions avec des indel connus (1000g, dnSNP) et détectés sont ciblées.**
  - **Pour une région donnée, trouve la meilleur séquence consensus.**
  - **Le score de la séquence consensus est égal à la somme des qualités des bases ayant un mésappariement**
  - **Si le score du consensus est supérieur à celui de l'alignement de départ, le nouvel alignement est conservé**

Analyse secondaire : Recalibration du score de qualité

- **Ce que dit le GATK :**
  - **Les constructeurs des séquenceurs actuels ont tendance à un peu trop sur-estimer les scores de qualité.**
  - **En général les séquenceurs actuels ont du mal dans les homopolymères**

Analyse secondaire : Recalibration du score de qualité

- **Comment ça marche :**
  - **Recherche des mésappariements dans l'alignement**
  - **Si un mésappariement n'est pas un variant connu alors on considère que c'est une erreur de séquençage**
  - **Calcul de covariance sur plusieurs caractéristiques (qualité des bases, position, cycle, contexte, etc …)**
  - **Correction du score de qualité**

Analyse secondaire : Détection des variants

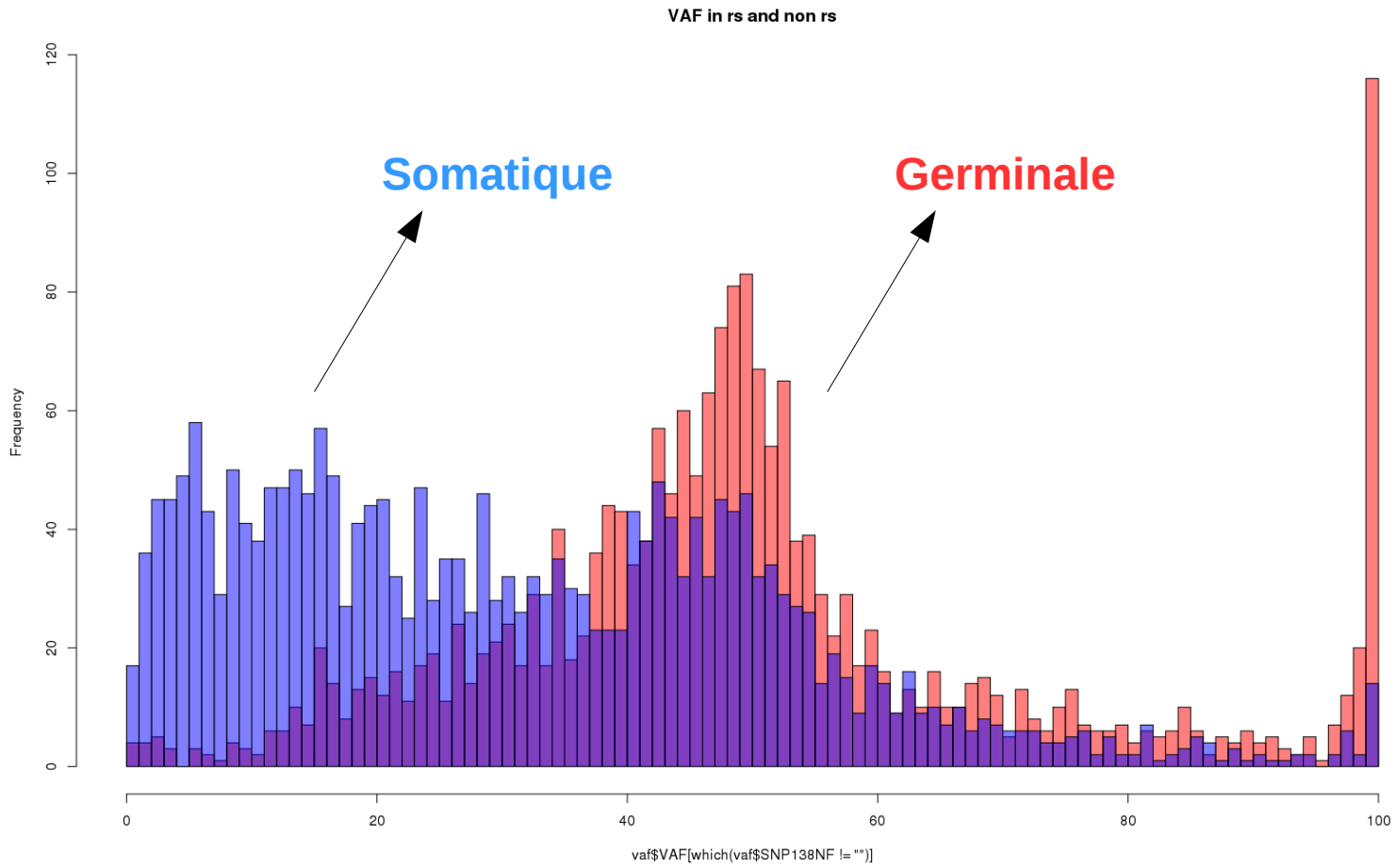
Name	OS	BAM/SAM input	Other inputs	Output	Identifies	Data set	Result[a]
Germline callers							
CRISP	Lin	Yes	–	VCF	SNP, INDEL	KTS	24 034 SNPs, 259 INDELs
GATK (UnifiedGenotyper)	Lin	Yes	–	VCF	SNP, INDEL	KTS	49 476 SNPs, 1959 INDELs
SAMtools	Lin	Yes	FASTA	VCF	SNP, INDEL	KTS	21 852 SNPs, 332 INDELs
SNVer	Lin, Mac, Win	Yes	–	VCF	SNP, INDEL	KTS	22 105 SNPs, 234 INDELs
VarScan 2	Lin, Mac, Win	No	pileup/mpileup	VCF, VarScan CSV	SNP, INDEL	KTS	34 984 SNPs, 1896 INDELs
Somatic callers							
GATK (SomaticIndelDetector)	Lin	Yes	–	VCF	INDEL	WES	151 INDELs
SAMtools	Lin	Yes	FASTA	BCF	SNP, INDEL	WES	Canceled[b]
SomaticSniper	Lin	Yes	–	VCF, somatic sniper output	SNP, INDEL	WES	6926 SNPs
VarScan 2	Lin, Mac, Win	No	pileup/mpileup	VCF, VarScan CSV	SNP, INDEL, CNV	WES	1685 SNPs, 324 INDELs
CNV identification tools							
CNVnator	Lin	Yes	FASTA	CSV	CNV	cnv.sim	39 CNVs
RDXplorer	Lin, Mac	Yes	FASTA	CSV	CNV	cnv.sim	4 CNVs[c]
CONTRA	Lin, Mac	Yes	FASTA	VCF, CSV	CNV	WES	3 CNVs
ExomeCNV	Lin, Mac, Win	Yes	pileup + BED + FASTA	CSV	CNV, LOH	WES	137 CNVs
SV identification tools							
BreakDancer	Lin, Mac	Yes	config file	CSV, BED	INDEL, INV, TRANS, CNV	WGS (tumor + normal)	6219 DELs, 0 INSs, 7 INVs, 17 303 ITX, 5037 CTX
Breakpointer	Lin	Yes	–	GFF	INDEL	WGS (tumor)	[d]
CLEVER	Lin	Yes	FASTA	CLEVER format	INDEL	WGS (tumor)	[d]
GASVPro (GASVPro-HQ)	Lin, Mac	Yes	–	clusters file	INDEL, INV, TRANS	WGS (tumor)	2529 DELs, 207 INVs
SVMerge	Lin	Yes	FASTA	BED	INDEL, INV, CNV	–	Aborted[e]

Stephan et al. A survey of tools for variant analysis of next-generation genome sequencing data Brief Bioinform 2013

Somatique Versus Germinale

- **Germinale :**
  - **Héritable**
  - **Présent dans toutes les cellules de la descendance**
  - **Cancer Familiaux (BRCA1)**
- **Somatique :**
  - **Non Héritable**
  - **Uniquement présent dans le clone de cellules portant la mutation**

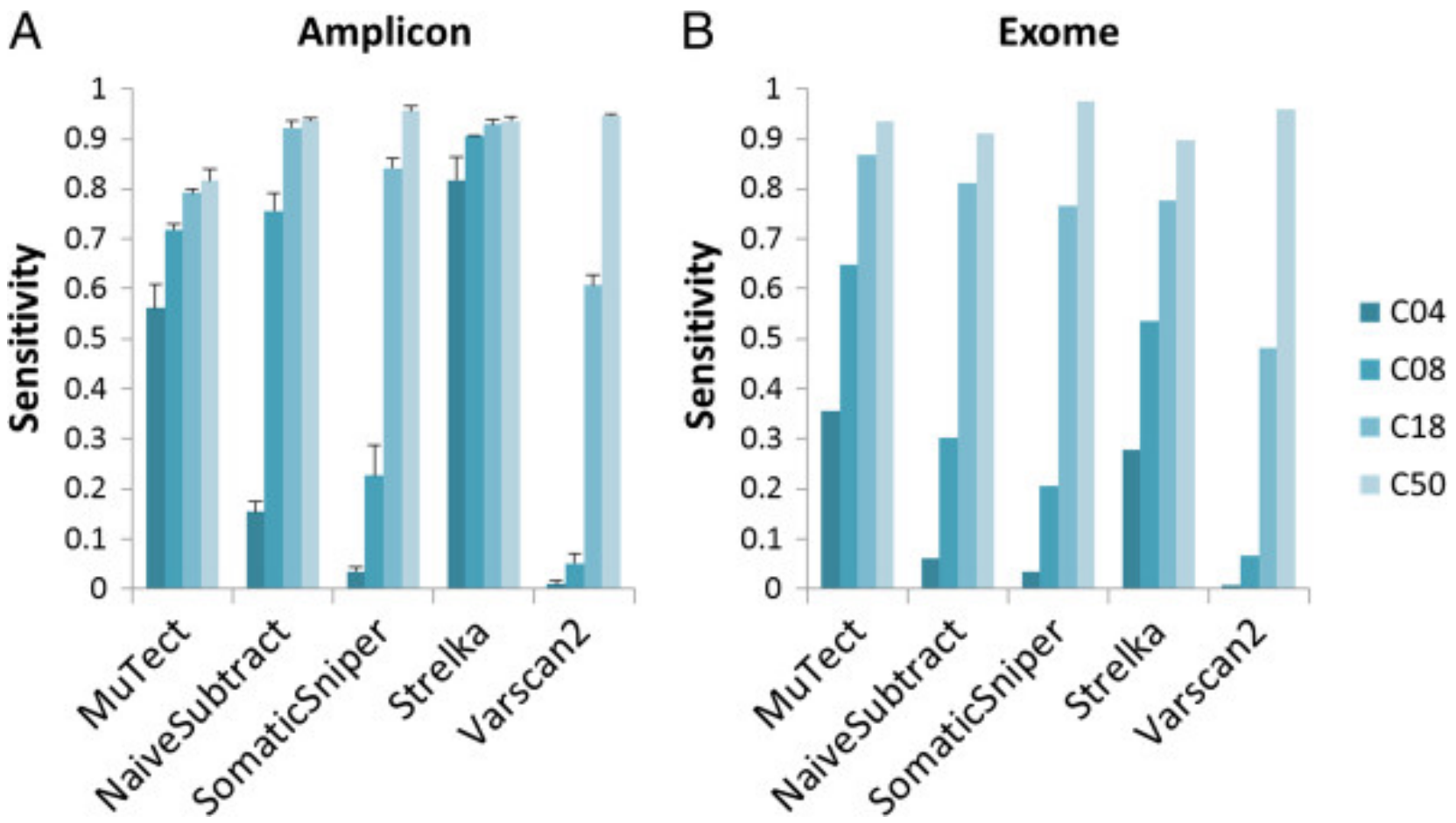
Somatique Versus Germinale



Mutation somatique : Détection

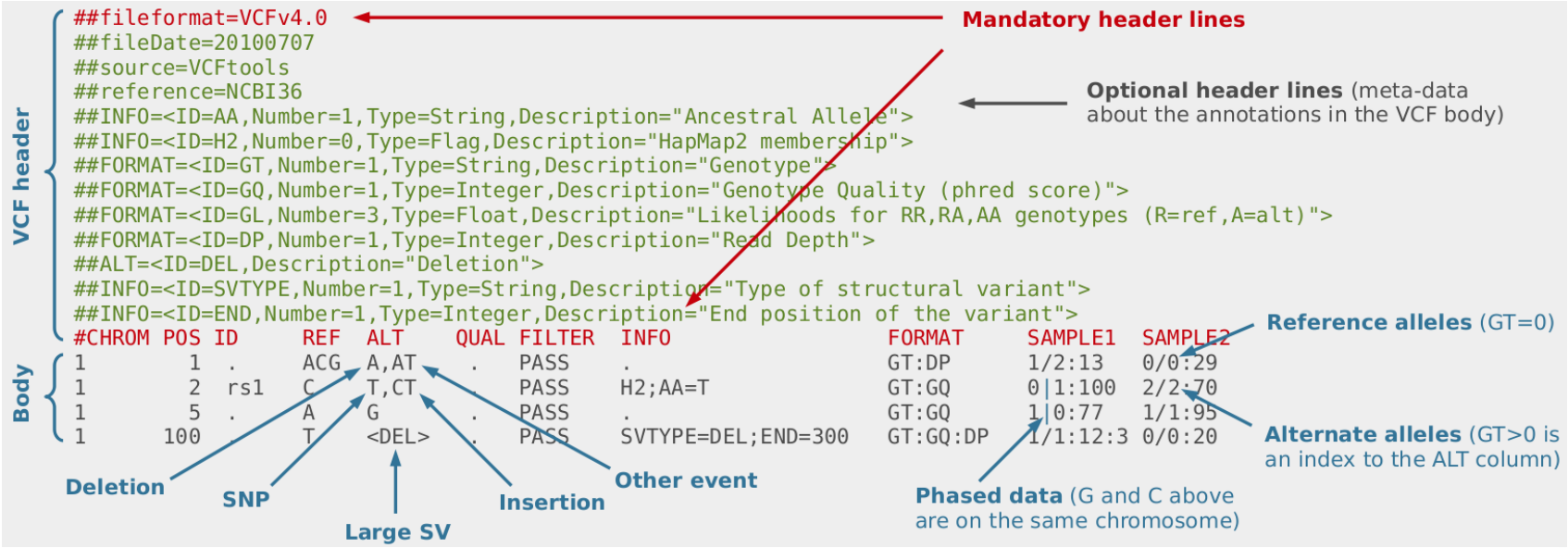
- Les mutations somatiques peuvent apparaître à des fréquences très faibles (< 5%) en partie à cause de :
  - L'hétérogénéité tumorale
  - La céllularité tumorale du prélèvement
- Requiert un niveau de sensibilité beaucoup plus important que pour la détection germinale
- Compromis Sensibilité/Spécificité compliqué
- Beaucoup d'outils disponible avec des performances variables.

Mutation somatique : Performances



Huilei Xu et al. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC Genomics2014

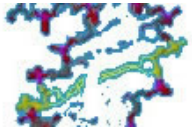
Format : VCF



- **Format standardisé pour enregistrer les variants**
- **Peut être indexé, compressé**
- **Flexible**

Analyse Tertiaire : L'annotation

dbSNP
Short Genetic Variations



ESP

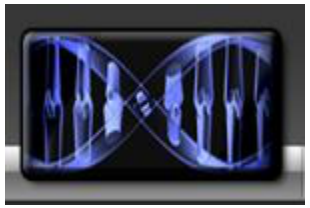
1000g

TATATCT
ACCTCAC
ClinVar

COSMIC
Catalogue of somatic mutations in cancer

UniProt

RefSeq



Polyphen

Analyse Tertiaire : L'annotation

Name	OS	Input	Output	SNP	INDEL	CNV	GUI	CLI	Web	Function/Location Parameters	DB IDs	Number of scores
ANNOVAR	Lin, Mac, Win, web interface	VCF, pileup, CompleteGenomics, GFF3-SOLiD, SOAPsnp, MAQ, CASAVA	TXT	Yes	Yes	Yes	No	Yes	No	9 (func) + 11 (exonic-func)	Yes	GERP++ conservation, LRT, MutationTaster, PhyloP conservation, PolyPhen, SIFT
AnnTools	Lin, Mac	VCF, pileup, TXT	VCF	Yes	Yes	Yes	No	Yes	No	5 (position) + 4 (functional class)	Yes	–
NGS–SNP	Lin, Mac	VCF, pileup, MAQ, diBayes, TXT	TXT	Yes	No	No	No	Yes	No	17	Yes	Condel, PolyPhen, SIFT
SeattleSeq	web interface	VCF, MAQ, CASAVA, GATK BED, custom	VCF, SeattleSeq	Yes	Yes	No	No	No	Yes	11 (dbSNP) + 5 (GVS)	Yes	GERP, Grantham, phastCons, PolyPhen
snpEff	Lin, Mac, Win	VCF, pileup/TXT (deprecated)	VCF, TXT, HTML overview	Yes	Yes	No	No	Yes	No	34	Yes	–
SVA	Lin	VCF, SV.events file, BCO	CSV	Yes	Yes	Yes	Yes	Yes	No	17 (SNP), 17 (INDEL), 10 (CNV)	Yes	–
VARIANT	web interface	VCF, GFF2, BED	web report, TXT	Yes	Yes	No	No	Yes	Yes	26	Yes	–
VEP	Lin, web interface	VCF, pileup, HGVS, TXT, variant identifiers	TXT	Yes	Yes	No	No	Yes	Limited	28	Yes	Condel, PolyPhen, SIFT

L'annotation : dbSNP

RefSNP	
Organism:	human (Homo sapiens)
Molecule Type:	Genomic
Created/Updated in build:	36/146
Map to Genome Build:	107/Weight
Validation Status:	

Allele	
Variation Class:	SNV: single nucleotide variation
RefSNP Alleles:	A/G (REV)
Allele Origin:	
Ancestral Allele:	A
Variation Viewer:	VarView
Clinical Significance:	NA
MAF/MinorAlleleCount:	T=0.4846/2427
MAF Source:	1000 Genomes

HGVS Names	Links
NC_000007.13:g.11584142T>C	
NC_000007.14:g.11544515T>C	
NG_027670.1:g.292683A>G	
NM_015204.2:c.1454-1398A>G	
XM_006715659.1:c.1454-1398A>G	
XM_006715660.1:c.1454-1398A>G	
XM_006715661.2:c.1454-1398A>G	
XM_006715662.1:c.1454-1398A>G	
XM_011515193.1:c.1247-1398A>G	

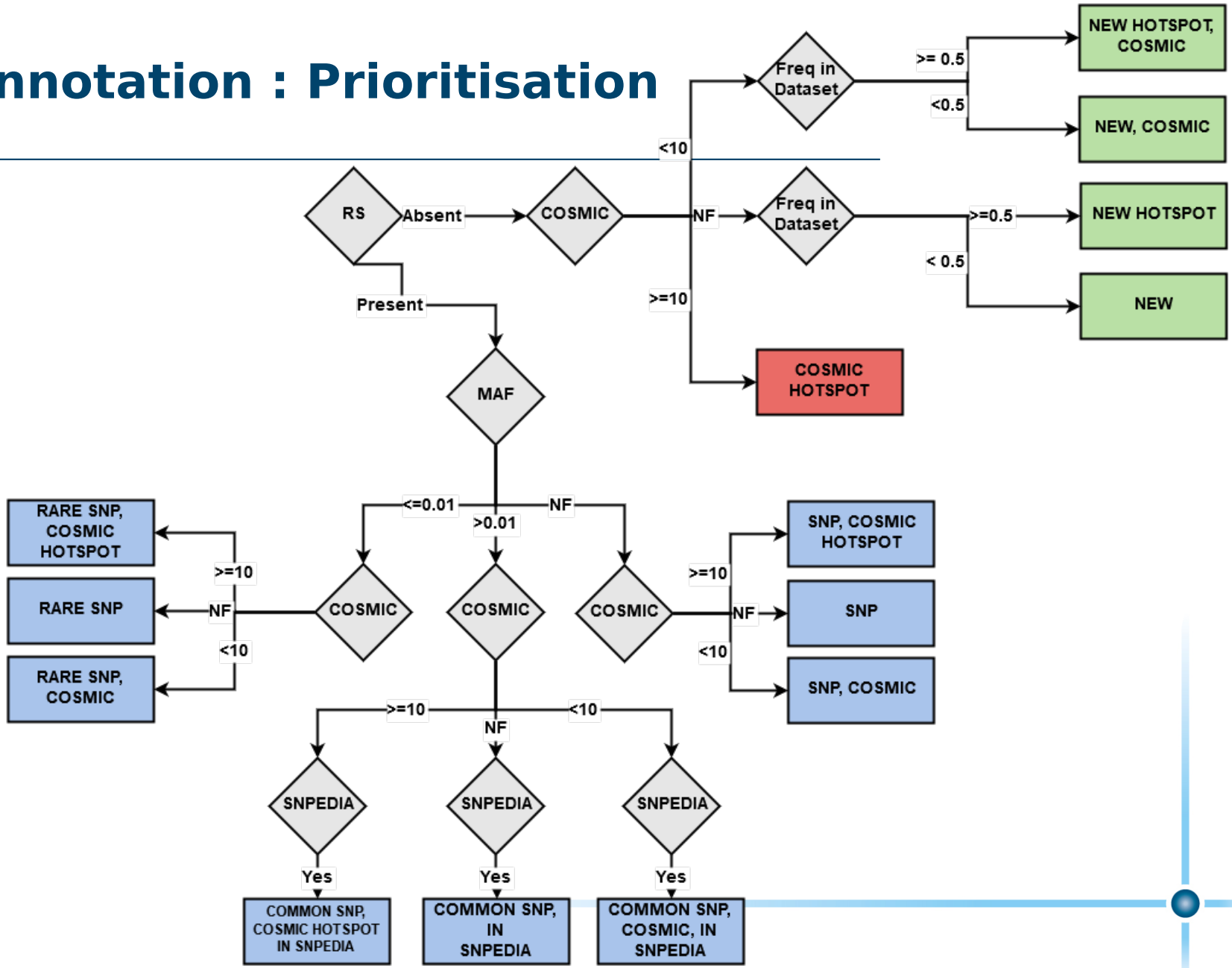
...more

	Sample Ascertainment				Genotype Detail				Alleles	
ss#	Population	Individual Group	Chrom. Sample Cnt.	Source	A/A	A/G	G/G	HWP	A	G
ss115929711	YRI		2	IG		1.00000000			0.50000000	0.50000000
ss1323907305	EAS		1008	AF					0.47419998	0.52579999
	EUR		1006	AF					0.52090001	0.47910002
	AFR		1322	AF					0.49319997	0.50680000
	AMR		694	AF					0.38329998	0.61669999
	SAS		978	AF					0.51840001	0.48159999

L'annotation : Cosmic

Tissue	Point Mutations		Copy Number Variation		Gene Expression		Methylation	
	% Mutated	Tested	Variant %	Tested	% Regulated	Tested	% Diff. Methylated	Tested
Adrenal gland		395		72		79		-
Autonomic ganglia		1109		-		-		-
Biliary tract		811		-		-		-
Bone		1006		-		-		-
Breast		11706		997		1104		-
Central nervous system		3736		811		695		-
Cervix		688		174		307		-
Endometrium		3753		424		602		-
Eye		168		-		-		-
Fallopian tube		4		-		-		-
Gastrointestinal tract (site indeterminate)		957		-		-		-
Genital tract		28		-		-		-

L'annotation : Prioritisation



Workflow : Ressources

- Le Base calling nécessite peu de ressource CPU/RAM
- L'alignement requiert beaucoup de RAM/CPU
- L'analyse bio-info d'un échantillon nécessite entre 2 et 24 heures
- Quantité de données générée très élevée (~Gb/Sample)
- Délais court pour rendre les résultats pour les projets de médecine personnalisée (PERMED/HEMATO BIO)

Traitement séquentiel sur machine personnel impossible

Solution : DISC

- Alambic
  - 236 CPU-CORES
  - 1168 Gb RAM
  - 18 noeuds
- Agui
  - 260 CPU-CORES
  - 1608 Gb RAM
  - 15 noeuds

Solution : Parallélisation

- **Parallélisation par les données**
  - **Simple à mettre en œuvre**
  - **Parfaitement adaptée à un environnement de type cluster de calcul (oar)**
  - **Les données doivent pouvoir être « découpées »**
  - **Les données doivent pouvoir être rassemblées**
  - **Limité par les processus I/O**

```
for i in $(ls $fastq_dir/*R1*)
do
   oarsub ...
done
```

Solution : Parallélisation

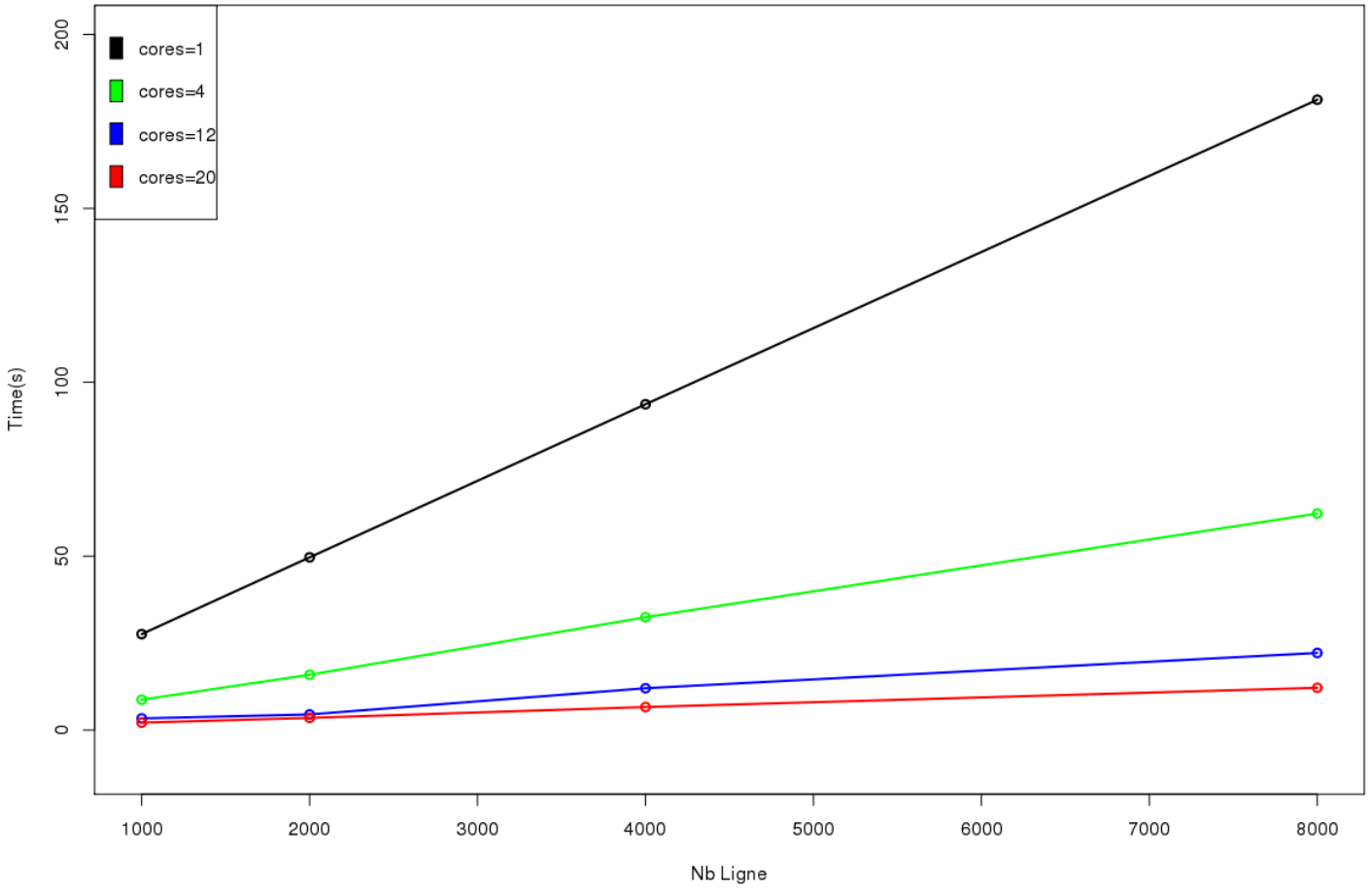
```
import multiprocessing

def worker():
    """worker function"""
    print 'Worker'
    return

if __name__ == '__main__':
    jobs = []
    for i in range(5):
        p = multiprocessing.Process(target=worker)
        jobs.append(p)
        p.start()
```

- Parallélisation par le code (Python : Multiprocessing)
  - Natif avec Python 2.6
  - Nécessite des compétences avancées en programmation

Solution : Parallélisation



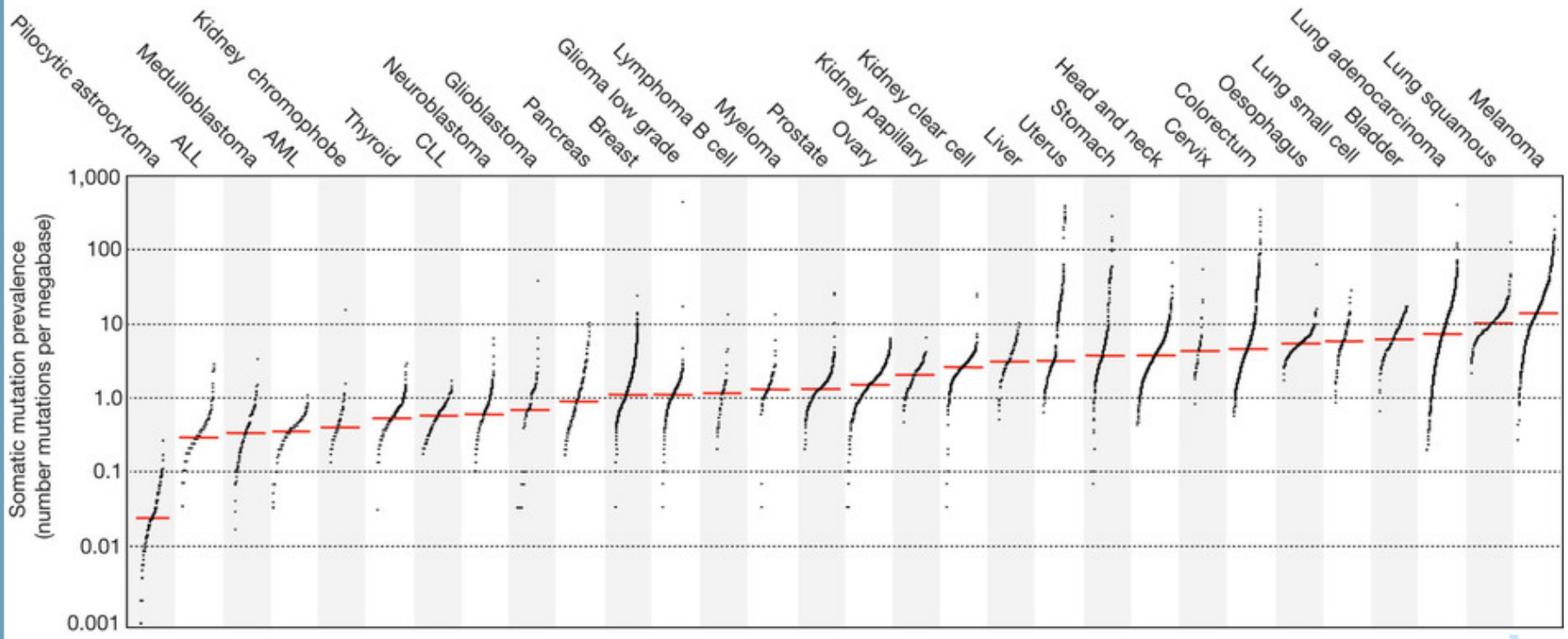
Résultats : Séquençage ciblé PERMED

Mutations HOTSPOT ou potentiellement dommageables :

Symbol	CNA	Mutation	Note	Freq	Drugs*
ERBB2	Normal	p.L755S	HotSpot	0.14	ado-trastuzumab.emtansine, afatinib, canertinib, ertumaxomab, lapatinib, neratinib, pelitinib, pertuzumab, sapitinib, trastuzumab, varlitinib, mubritinib, margetuximab, ado-trastuzumab.emtansine, .T-DM1
ESR1	Normal	p.D538G	Damaging	0.13	Tamoxifen Anti-A
MLL2	Normal	p.G1995R	Damaging	0.31	
RAD51D	Not found	p.S144F	Damaging	0.52	
RB1	Normal	p.R621C	Damaging	0.56	
SACS	Normal	p.D4116Y	Damaging	0.24	
STAT5B	Normal	p.V696I	Damaging	0.15	dasatinib
USH2A	Normal	p.G1871D	Damaging	0.69	
WRN	Loss or Del	p.T1262R	Damaging	0.56	

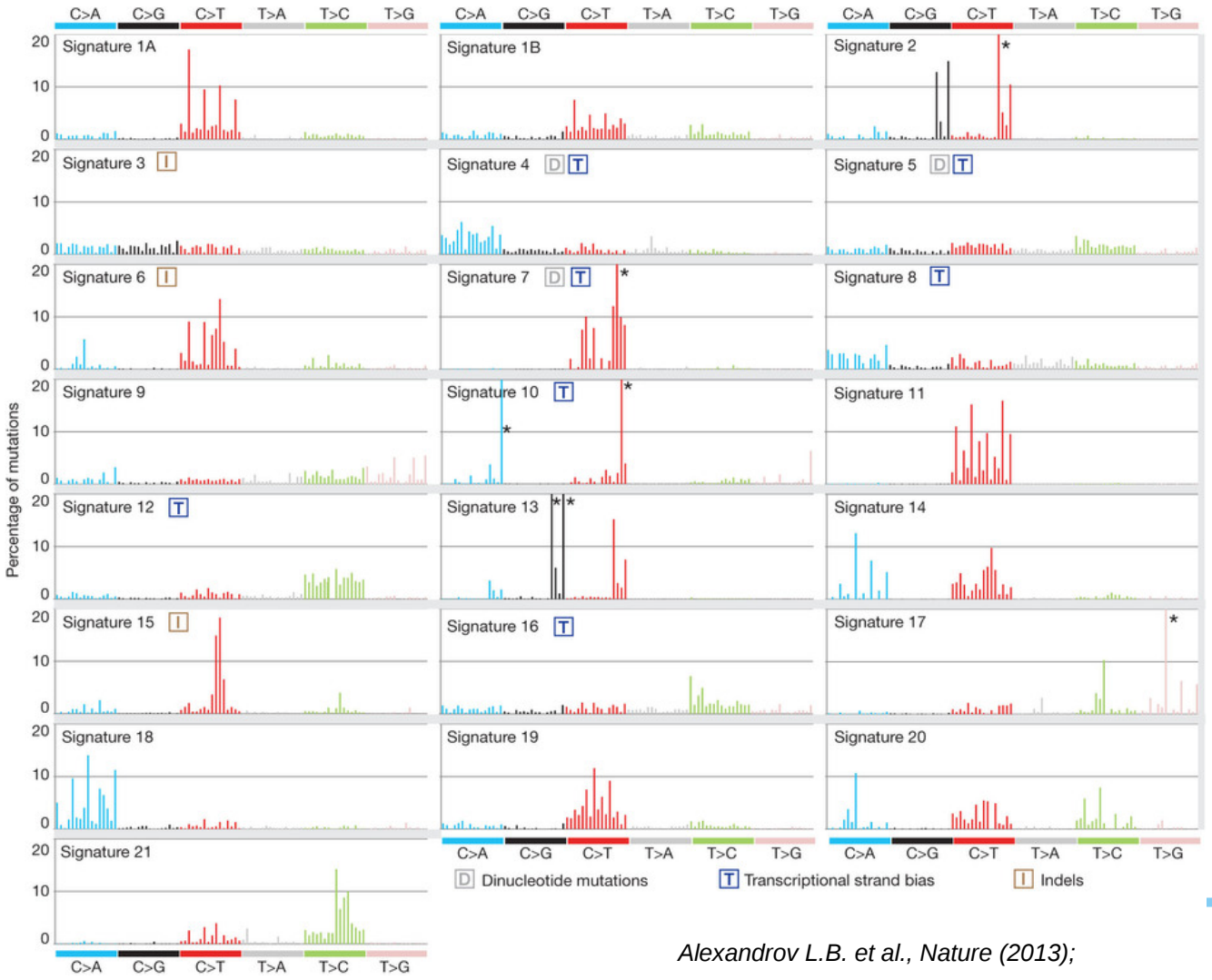
* Cliquez sur le gène pour plus de détails

Résultats : Signature mutationnelle



Alexandrov L.B. et al., Nature (2013);

Résultats : Signature mutationnelle



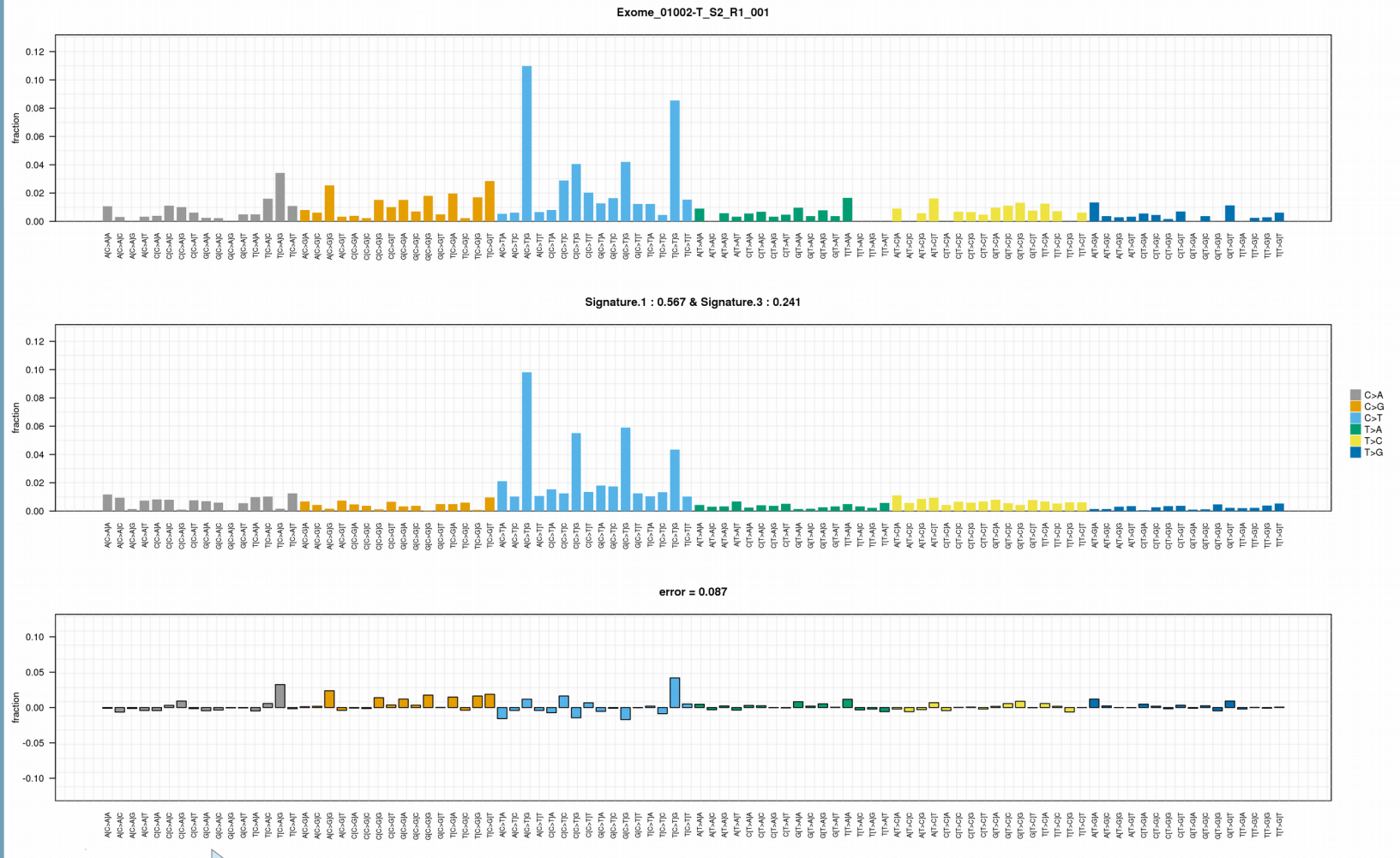
Alexandrov L.B. et al., Nature (2013);

Résultats : PERMED 1002

Symbol	CNA	Loc	Mutation	Note	Freq
ADAMTS16	Normal	chr5 :5182400	NM_139056 exon4 c.T745G p.C249G	DAMAGING	0.41
BRCA1	Normal	chr17 :41256250	NM_007294 exon6 c.G330T p.K110N	DAMAGING	0.70
MAP3K4	Normal	chr6 :161470401	NM_005922 exon3 c.C1097G p.A366G	DAMAGING	0.61
MYH9	Normal	chr22 :36681312	NM_002473 exon38 c.C5338T p.R1780W	DAMAGING	0.88
NTRK1	Normal	chr1 :156846234	NM_002529 exon14 c.C1675T p.R559C	DAMAGING	0.02
PPP2R1A	Normal	chr19 :52714679	NM_014225 exon4 c.C437T p.S146L	DAMAGING	0.43
PTEN	Normal	chr10 :89692904	NM_000314 exon5 c.C388G p.R130G	HOTSPOT	0.53
PTEN	Normal	chr10 :89685270	NM_000314 exon3 c.166dupT p.R55...	frameshift insertion	0.35
SMARCA4	Normal	chr19 :11094931	NM_003072 exon2 c.C104T p.S35L	DAMAGING	0.86
STAT5B	Normal	chr17 :40353869	NM_012448 exon19 c.C2251T p.L751F	DAMAGING	0.45
TP53	Normal	chr17 :7577599	NM_000546 exon7 c.681dupT p.D22...	stopgain SNV	0.57

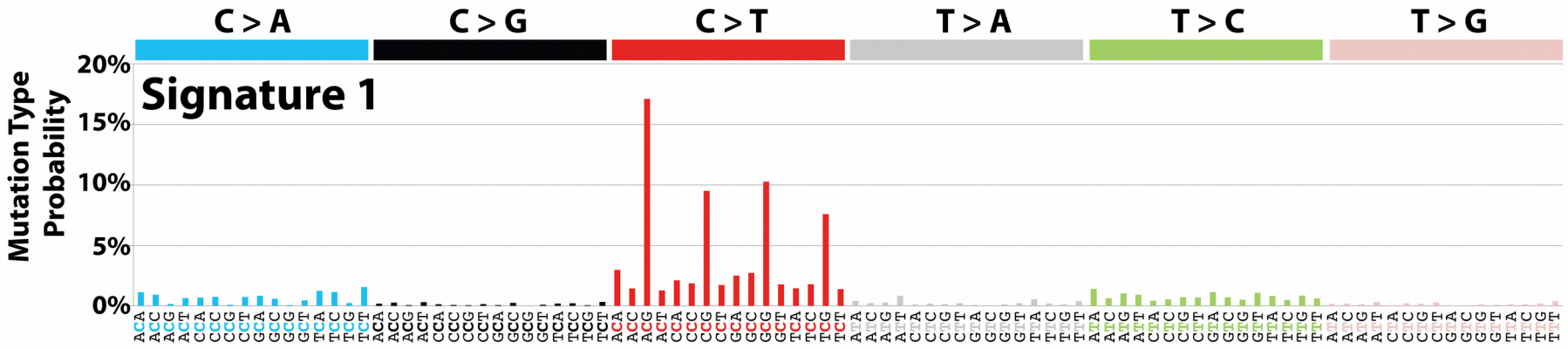
* Cliquez sur le gène pour plus de détails

Résultats : Exome PERMED 1002



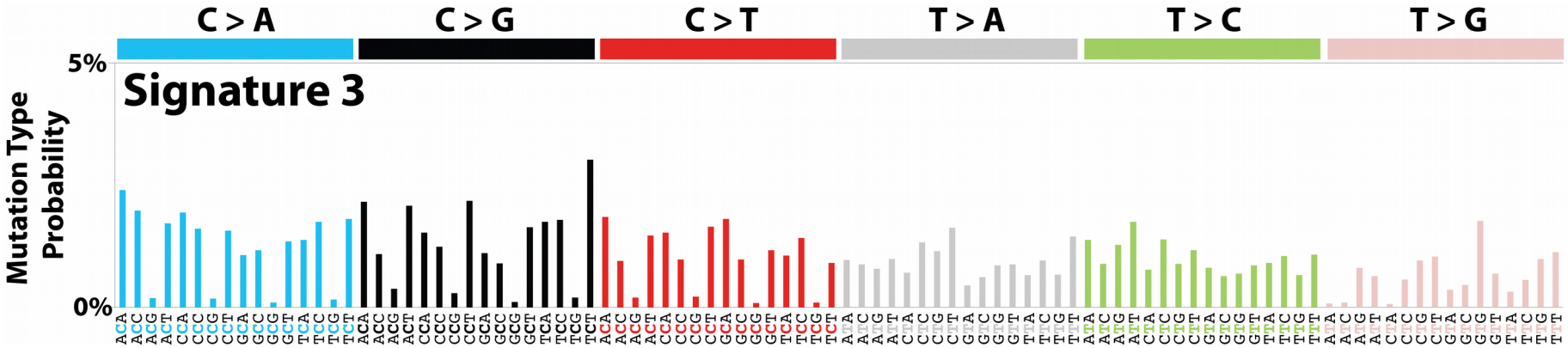
PERMED 1002 associée avec la signature 1 et 3

Résultats : Exome PERMED 1002



- **Types de cancer : la signature 1 a été trouvée dans presque tous les types de cancer.**
- **Étiologie : la signature 1 est le résultat d'un processus mutationel endogène initié par la désamination des 5-methylcytosines**
- **Caractéristiques des mutations : la signature 1 est associée à un petit nombre de petits indels dans la plupart des tissus.**

Résultats : Exome PERMED 1002



- **Types de cancer : trouvée dans les cancers du sein, de l'ovaire et du pancréas.**
- **Étiologie : associée à un défaut de la réparation par recombinaison homologue**
- **Caractéristiques des mutations : fortement associée à un grand nombre de large indel (>3bp)**
- **Commentaire : fortement associée à des mutations germinales ou somatiques dans BRCA1 et/ou BRCA2. Dans le pancréas les patients sensibles au platine présentent cette signature.**

Merci !!!

Des questions ???