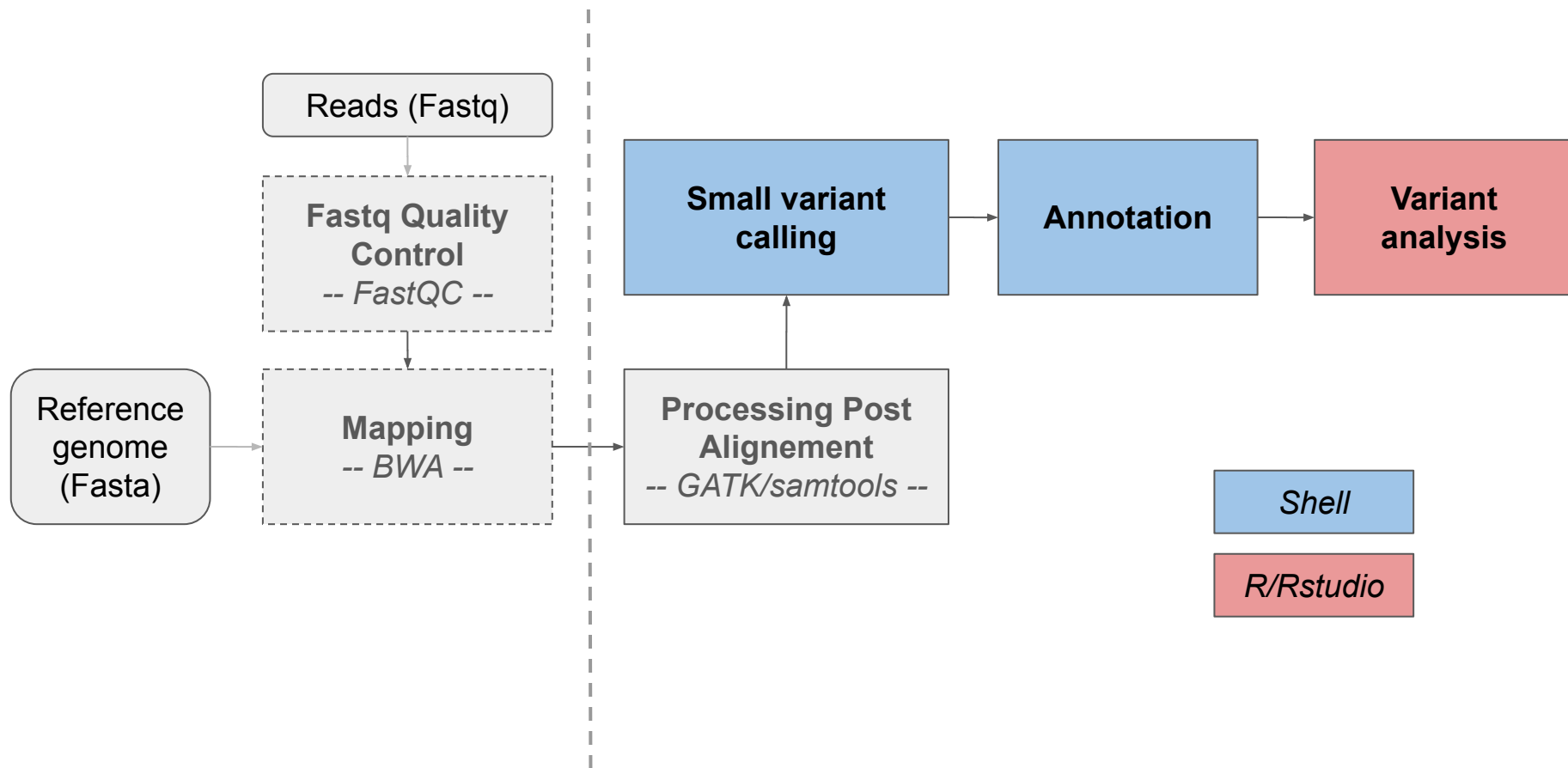




Variant calling

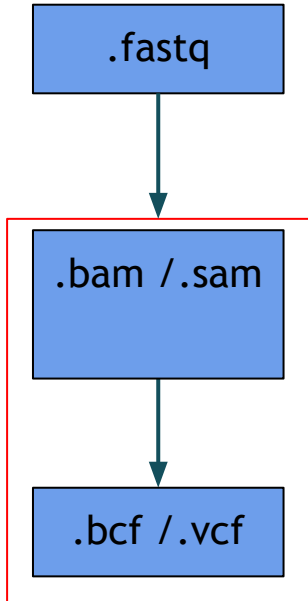
Vivien Deshaies - AP-HP

Workflow



Qu'appelle t-on "Variant Calling"

Détection automatisée des variants (SNVs, Indels de petite taille) à partir d'un fichier contenant des données de séquençage alignées (BAM)



```
@H5:1:H3T27BBXY:8:1101:1955:1191/1
ATTNTTATAGATTCTAGGAAGTTGCTCGAGAAGTTTTCTAATTAGTAGAAGTTGTTGGAGAAGCGTCTAGTTAGCGGAAGTAGCTCGAGAAGCTTCTATT CAGTAATATATATAAGAGTCGAGG
+
AAA#FJJFJJJJFFJJJJJJJJFJJFJJJJJJ<<AJJJJJJJJJJJJA<JJFJJJJJJJJJJFF<<JJJFJJJJFAJFJJ<JFJJJJJJJJF<FJJAJ-FFJFFAAAFJ<A-FJFJJ-7FFFJ
```

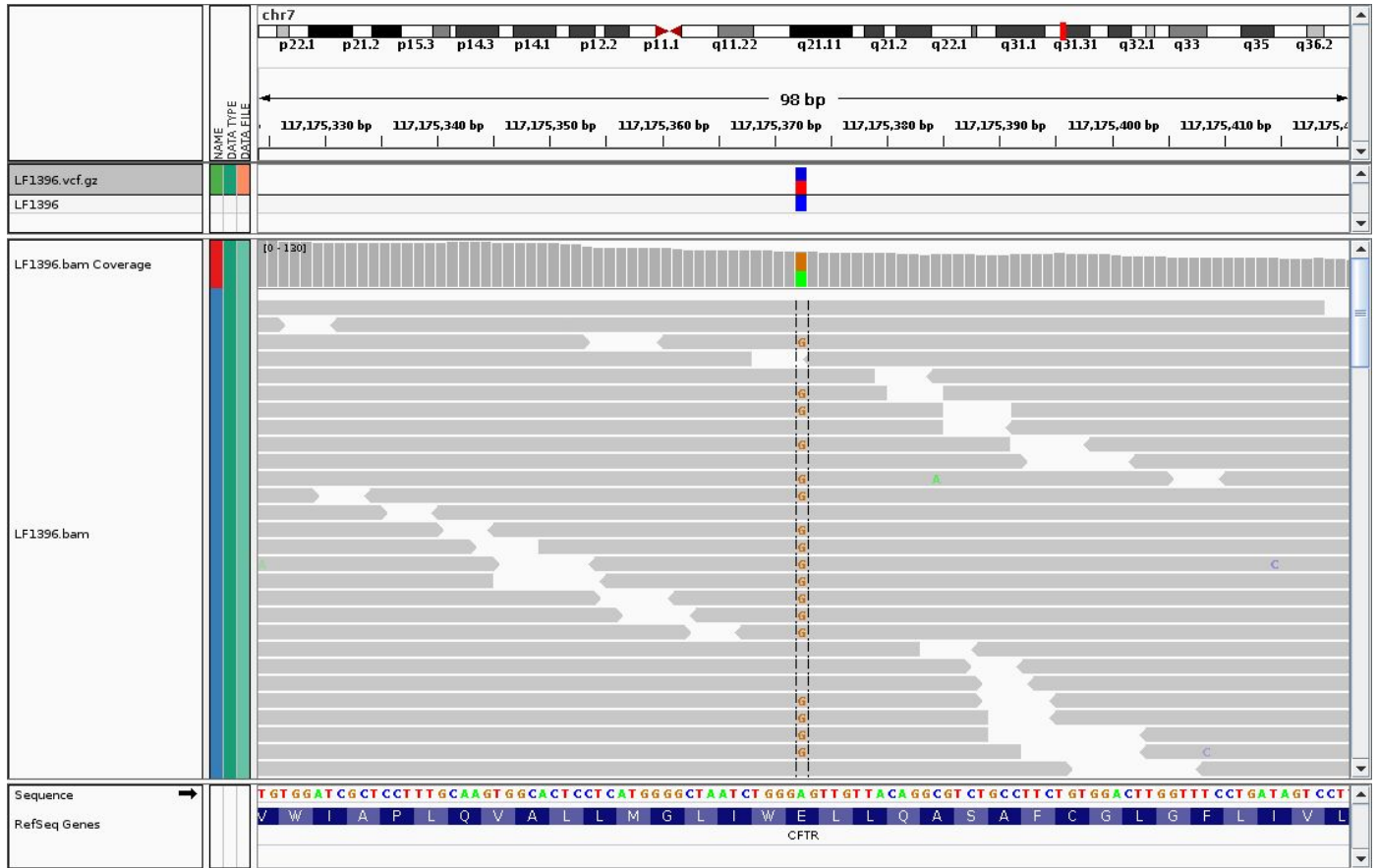
H5:1:H3T27BBXY:8:1110:4878:2035	83	Chr01	1568	60	136M	=	1495	-209	AAACCCTAACCCCTAACCCCTAACCCCTAA
H5:1:H3T27BBXY:8:1128:11657:35198		99	Chr01	1572	60	151M	=	1843	422 CCTAACCCCTAACCCCTAACCCCTAACCC
H5:1:H3T27BBXY:8:1217:6045:36200		163	Chr01	1575	60	115M	=	1575	126 AAACCCTAACCCCTAACCCCTAACCCCTAA
H5:1:H3T27BBXY:8:1217:6045:36200		83	Chr01	1575	60	126M	=	1575	-126 AAACCCTAACCCCTAACCCCTAACCCCTAA
H5:1:H3T27BBXY:8:2227:16863:39963		83	Chr01	1582	60	89M	=	1560	-111 AACCCCTAACCCCTAACCCCTAACCCCTAA

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Ech	-456
Chr2	1091	.	C	A	161.77	.	AC=1;AF=0.500;AN=2;BaseQRankSum=0.672;ClippingRankSum=0.567;DP=44;Exces			
Chr2	1226	.	T	A	618.77	.	AC=1;AF=0.500;AN=2;BaseQRankSum=-6.233;ClippingRankSum=1.014;DP=201;Ex			
Chr2	1708	.	G	A	133.77	.	AC=1;AF=0.500;AN=2;BaseQRankSum=0.000;ClippingRankSum=-0.720;DP=6;Exces			

Qu'appelle t-on "Variant Calling"

.vcf

.bam / .sam

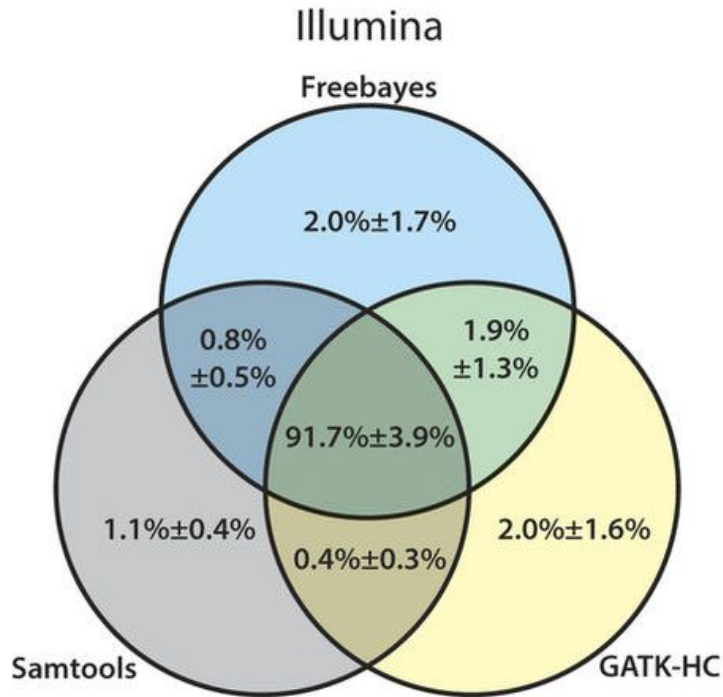


Variant callers

- Choix du variant caller en fonction de la question biologique
- Utilisés classiquement par la communauté :
 - GATK Haplotype Caller
 - Samtools mpileup/Bcftools
 - Samtools mpileup/VarScan2
 - FreeBayes
 - GATK Mutect2 (spécifique à la détection tumorale)
 - DiscoSnp (variant calling sans génome de référence)
 - DeepVariant : variant calling par analyse d'images pileup (regions complexes, low depth)

→ **Aucun outil n'est parfait** : la qualité du calling dépend de l'ensemble du pipeline, des données analysées, et des paramètres utilisés pour filtrer les résultats

Concordance entre variant callers



- **Concordance de 91.7%** entre Freebayes, Samtools, GATK HC (Hwang et al., 2015)
- D'autres analyses montrent des taux plus bas :
 - **70%** (O'Rawe et al., Genome Med, 2013)
 - **57%** (Cornish et al., BioMed, 2015)
- La **sensibilité** et la **précision** diffèrent selon les outils et les paramètres utilisés

!/\\ Existence de variants qui sont spécifiques aux différents callers !/

Difficultés - Limitations

- De nombreux variants **Faux Positifs** peuvent survenir des étapes précédentes :
 - Artéfacts issus des **cycle PCR** pendant la préparation des échantillons
 - Artéfacts liés à la **technologie de séquençage** (PacBio, HiSeq, NextSeq, ...)
 - Difficultés d'**alignement** (régions d'ADN répétées)
 - **Erreurs de lecture** lors du “BaseCalling”
- Des algorithmes complexes de détection compliquent l'interprétation des résultats

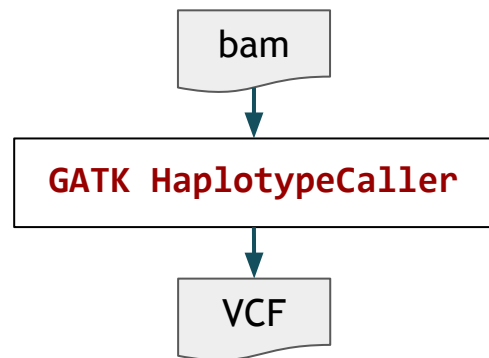
En conclusion

- La détection de variant permet d'identifier des SNVs et petits Indels à partir d'un fichier d'alignement au format BAM
- De nombreux outils existent pour la détection de variants, leur efficacité dépend de nombreux paramètres (mapping, qualité des données, paramètres de filtrage des résultats)
- La “sensibilité” et la “précision” permettent d'évaluer la qualité des résultats de détection de variant. Pour un même outil ces mesures varient selon les seuils de qualité utilisés.

Aller au `jupyterNoteBook`

1 / GATK HaplotypeCaller avec sortie VCF

Single-sample variant calling



VCF (variant call format)

VCF header

```
##fileformat=VCFv4.2
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filte...
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the ...
##GATKCommandLine=<ID=HaplotypeCaller,CommandLine="HaplotypeCaller --min-base-quality-score 18 --emit-ref-confidence NO...
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
...
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read posit...
##INFO=<ID=SOR,Number=1,Type=Float,Description="Symmetric Odds Ratio of 2x2 contingency table to detect strand bias">
##contig=<ID=6,length=119458736>
##source=HaplotypeCaller
```

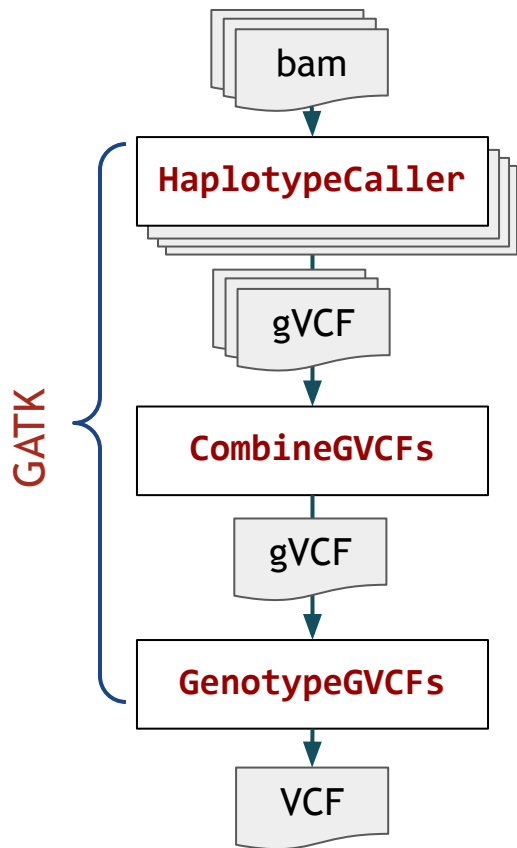
Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SRR1262731
6	37913396	.	T	A	67.64	.	AC=1;AF=0.500;...	GT:AD:DP:GQ:PL	0/1:3,2:5:75:75,0,105
6	37916445	.	GT	G	58.60	.	AC=1;AF=0.500;...	GT:AD:DP:GQ:PL	0/1:1,2:3:28:66,0,28
6	37921683	.	C	CA	55.60	.	AC=1;AF=0.500;...	GT:AD:DP:GQ:PL	0/1:7,2:9:63:63,0,279

SNP
Insertion
Deletion

2/GATK HaplotypeCaller en mode GVCF

Multi-sample variant calling



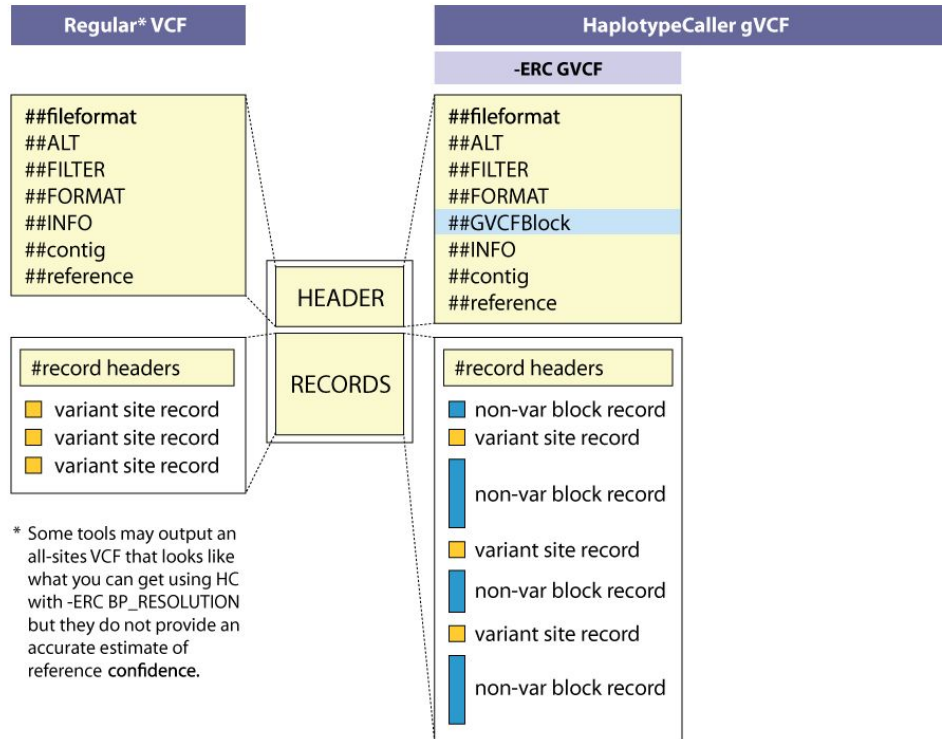
- En 3 étapes (=> 3 outils) :

Variant-calling avec sortie gvcf / par échantillon

Combiner les sorties gvcf en 1 sortie gvcf

Identifier les variants simultanément sur tous les échantillons

Sorties VCF vs. gVCF (option -ERC)



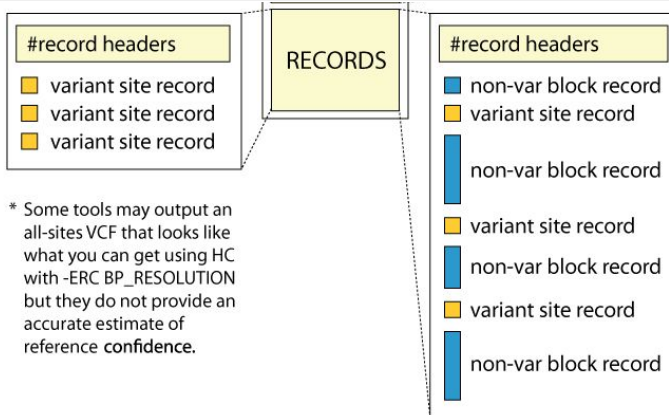
Sorties VCF vs. gVCF (option -ERC)

VCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SRR1262731
6	37913396	.	T	A	67.64	.	AC=1;AF=0.500;...	GT:AD:DP:GQ:PL	0/1:3,2:5:75:75,0,105
6	37916445	.	GT	G	58.60	.	AC=1;AF=0.500;...	GT:AD:DP:GQ:PL	0/1:1,2:3:28:66,0,28
6	37921683	.	C	CA	55.60	.	AC=1;AF=0.500;...	GT:AD:DP:GQ:PL	0/1:7,2:9:63:63,0,279

gVCF

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SRR1262731
6	37913111	.	G	<NON_REF>	.	.	END=37913131	GT:DP:GQ:MIN_DP:PL	0/0:3:9:3:0,9,114
6	37913132	.	A	<NON_REF>	.	.	END=37913133	GT:DP:GQ:MIN_DP:PL	0/0:4:12:4:0,12,170
...									
6	37913394	.	T	<NON_REF>	.	.	END=37913395	GT:DP:GQ:MIN_DP:PL	0/0:5:12:5:0,12,180
6	37913396	.	T	A,<NON_REF>	67.64	.	BaseQRankSum...	GT:AD:DP:GQ:PL:SB	0/1:3,2,0:5:75:75,...
6	37913397	.	A	<NON_REF>	.	.	END=37913400	GT:DP:GQ:MIN_DP:PL	0/0:5:12:5:0,12,180



VCF Multi-échantillons

VCF header

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Mandatory header lines (points to ##fileformat=VCFv4.0)

Optional header lines (meta-data about the annotations in the VCF body) (points to ##INFO=...)

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0/1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1/0:77	1/1:95
1	100	.	T		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Reference alleles (GT=0) (points to 0/0:29)

Alternate alleles (GT>0 is an index to the ALT column) (points to 2/2:70)

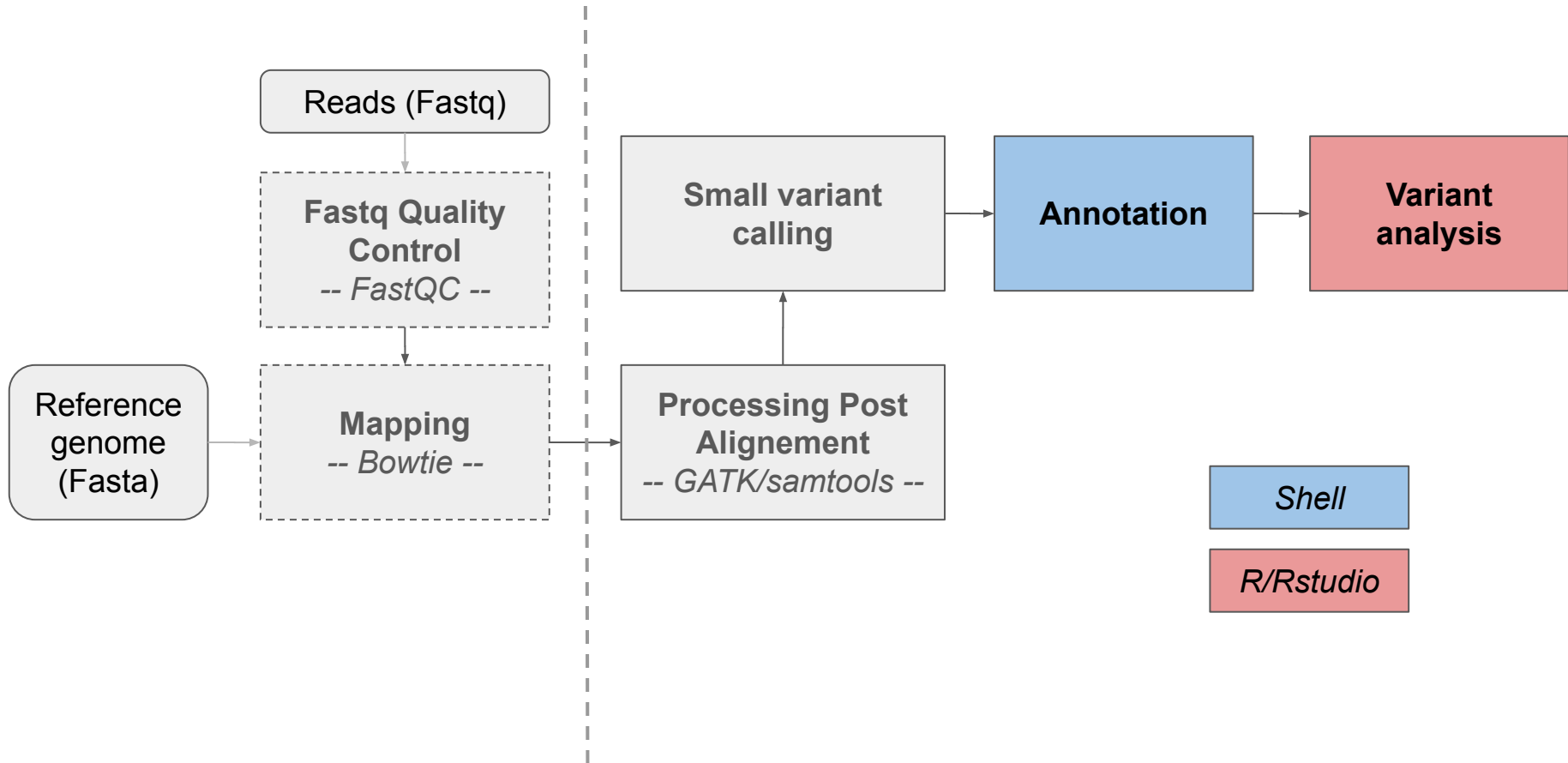
Deletion (points to in ALT)

SNP (points to A,AT in ALT)

Insertion (points to T,CT in ALT)

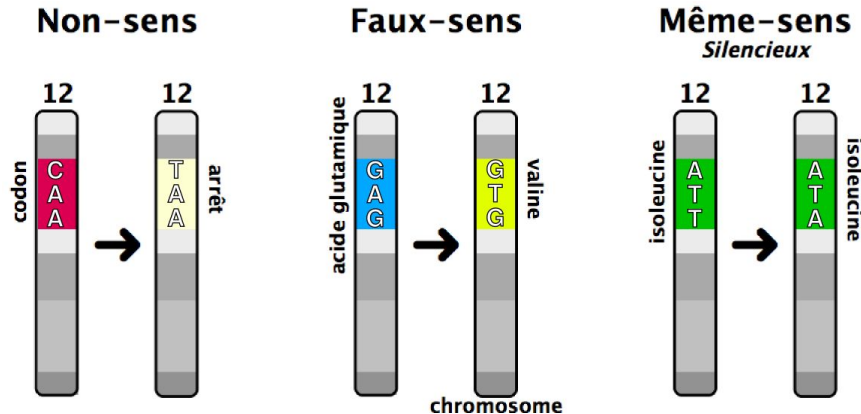
Other event (points to H2;AA=T in INFO)

Workflow



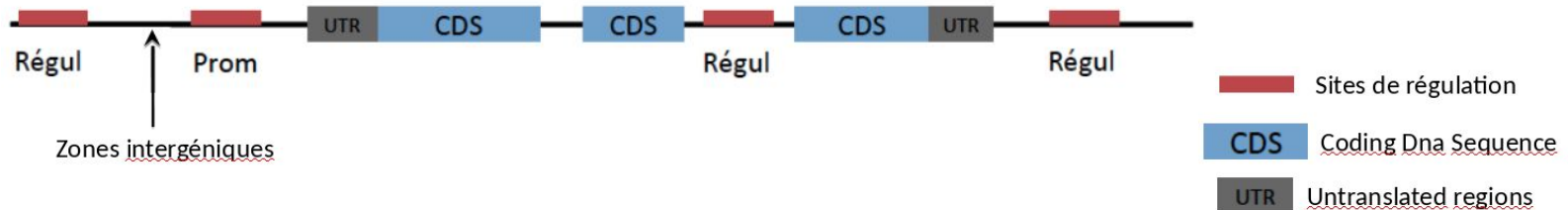
Annotation des variants

- Ajout d'**informations biologiques pertinentes** aux variants :
 - Est-ce que mes variants sont connus ?
 - Où se positionnent mes variants ?
 - Quel est l'effet d'une mutation sur le CDS qui le contient ?



Annotation des variants

- Annotation structurale :
→ Mon variant se trouve-t-il dans un **intron**, un **exon** ?
- Annotation fonctionnelle :
→ Informations sur la région ? Exemple : CDS codant pour une protéine
- Impacts potentiels :
→ Dans le cas d'un CDS, **protéine produite tronquée**, allongée, décalée... ou silencieuse (redondance du code génétique)



Annotation des variants

- Nécessité d'avoir des **bases de données** associées aux organismes étudiés (Ensembl, Refseq...)
- Exemples d'outils/algorithmes :
 - SnpEff
 - VEP
 - Annovar
 - SIFT, POLYPHEN2, CADD...
 - dbNSFP,

Aller au `jupyterNoteBook`