Atelier "Bioinformatique et analyse de données NGS"
IRD Ouagadougou, Burkina Faso
14/10/2016

Introduction aux analyses NGS

Du fichier fastq jusqu'au fichier de SNP

Christine Tranchant-Dubreuil - *Équipe RICE, UMR DIADE*

Alexis Dereeper - *UMR IPME*

Plateau i-trop, http://bioinfo.mpl.ird.fr

Les fichiers en début de TP

Répertoire Illumina :

- 2 fichiers RNA-seq au format fasq (illumina)

reçus de la société de séquençage

- un fichier contenant les séquences des adaptateurs utilisés

utile pour l'étape de "cleaning/nettoyage"

- Référence au format fasta

Utile pour l'étape de mapping

Quels logiciels utilisés?

- Contrôle de la qualité de séquençage *fastqc*
- "Nettoyage" des séquences : *cutadapt*
  - Retrait des adaptateurs
  - Remplacement des bases de mauvaise qualité (suite au séquençage) par des N (valeur PHRED > 30)
  - Retrait des extrémités des reads de mauvaise qualité (beaucoup de N)
  - Retrait des séquences trop petites suite aux étapes précédentes
- Mapping des séquences contre la référence sequence *bwa aln & sampe*
- Détection des SNP *samtools mpileup*

  bcftools

Quels logiciels utilisés?

- Contrôle de la qualité de séquençage

  fastq -> ***fastqc***

- “Nettoyage” des séquences :

  fastq -> ***cutadapt -> fastq***

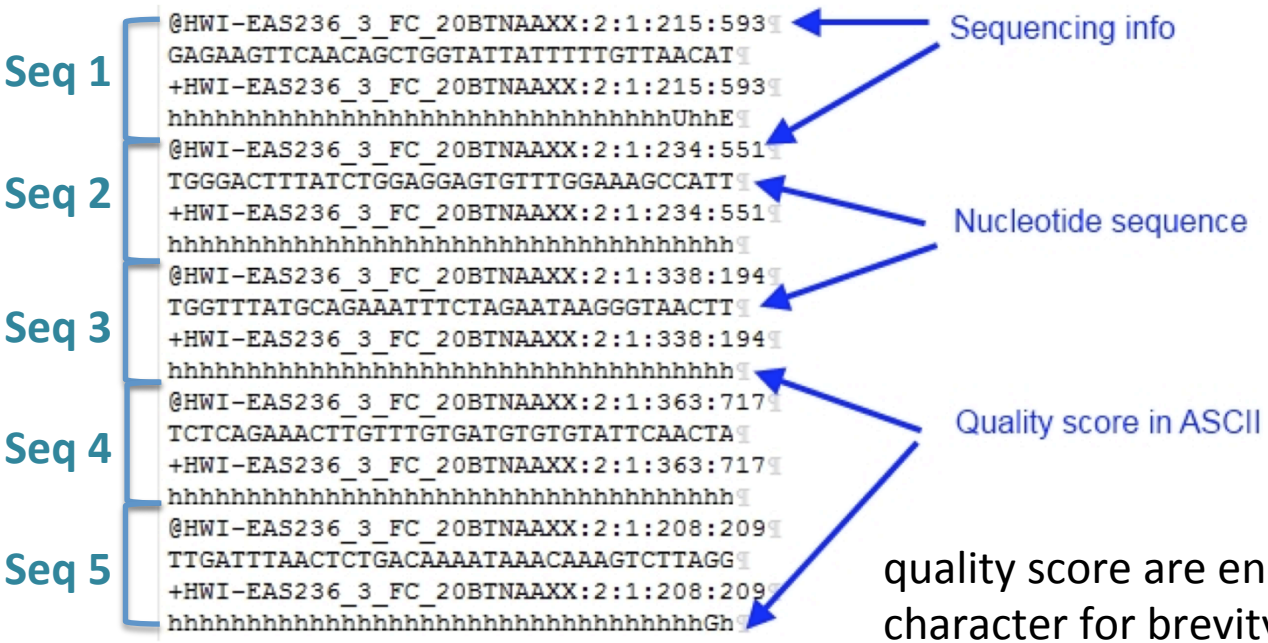
- Mapping des séquences contre la référence sequence

  fastq -> ***bwa aln*** + ***bwa sampe*** -> ***sam*** -> ***bam***

- Détection des SNP

  bam -> ***samtools mpileup***-> ***bcf*** -> ***bcftools*** -> ***vcf***

Format fastq



quality score are encoded with a single ASCII character for brevity

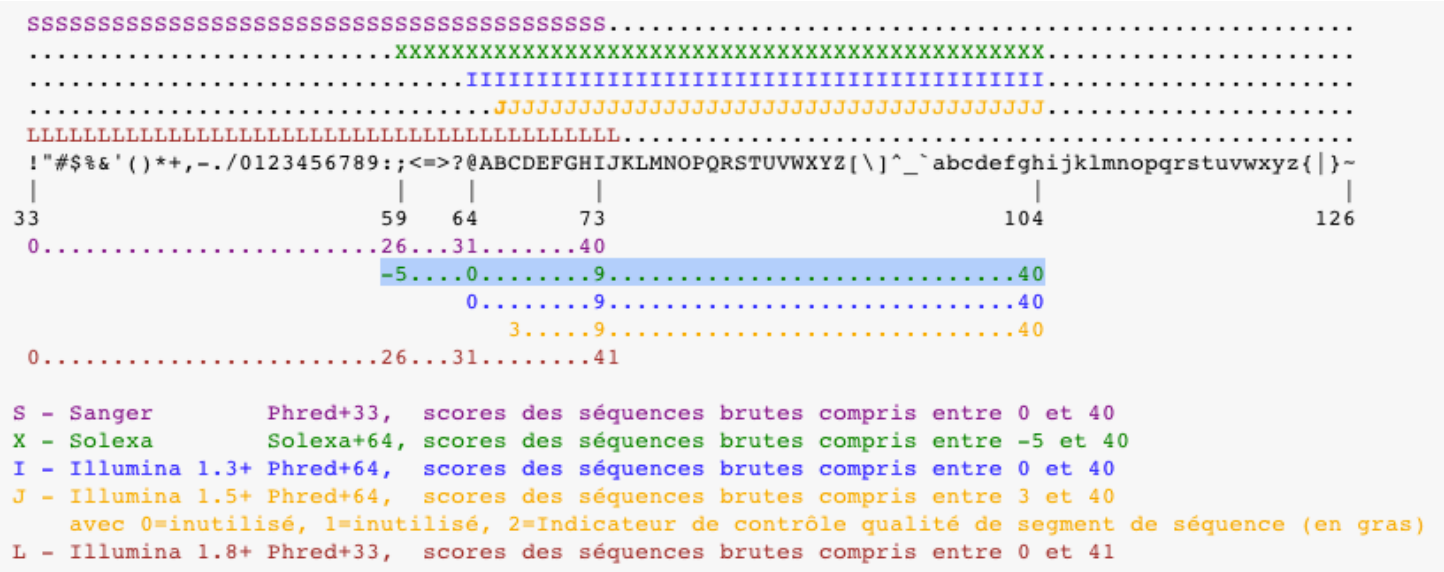
1 séquence = 4 lignes

- @ identifiant séquence
- sequence
- + nom séquence name (optionel).
- Qualité de la séquence (un caractère / base)

Qualité encodé par un seul caractère ASCII

Problème: différentes manières de coder la qualité de la séquence selon la technologie de séquençage illumina utilisée

Quality score encoded with a single ASCII character



Encodage Sanger code un score de qualité de Phred :

- de 0 to 40
- utilise ASCII 33 à 73

Quality sequencing

- Phred quality scores Q are defined as a property which is logarithmically related to the base-calling error probabilities P

$$Q = -10 \log_{10} P$$

or

$$P = 10^{\frac{-Q}{10}}$$

- For example: if Phred assigns a quality score of 30 to a base, the chances that this base is called incorrectly are 1 in 1000.

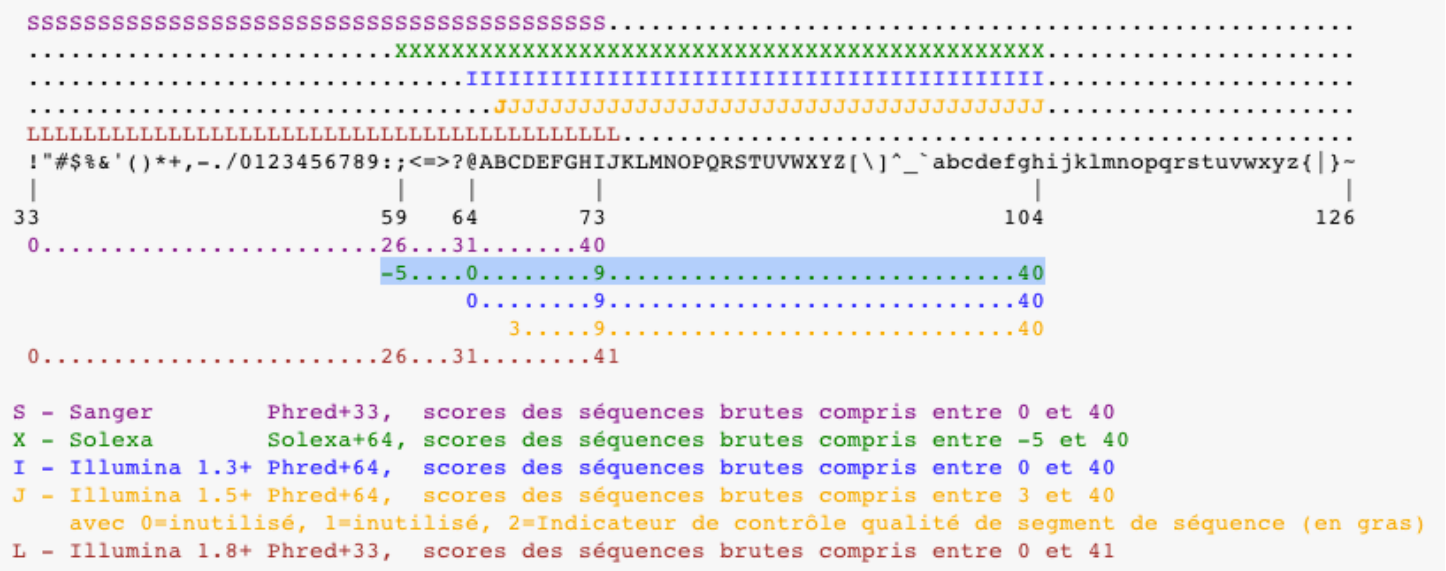
- The most commonly used method is to count the bases with a quality score of 20 and above.

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

Quality score encoded with a single ASCII character

Problème: différentes manières de coder la qualité de la séquence selon la technologie de séquençage illumina utilisée



Encodage Solexa/Illumina 1.3 / 1.5 format code un score qualité phred :

- de -5 à 40
- utilise ASCII 59 à 104

Quality score encoded with a single ASCII character

Problème: différentes manières de coder la qualité de la séquence selon la technologie de séquençage illumina utilisée

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ..............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                               0........9.............................40
                                  3.....9.............................40
 0........................26...31........41

S - Sanger        Phred+33,  scores des séquences brutes compris entre 0 et 40
X - Solexa        Solexa+64, scores des séquences brutes compris entre -5 et 40
I - Illumina 1.3+ Phred+64,  scores des séquences brutes compris entre 0 et 40
J - Illumina 1.5+ Phred+64,  scores des séquences brutes compris entre 3 et 40
    avec 0=inutilisé, 1=inutilisé, 2=Indicateur de contrôle qualité de segment de séquence (en gras)
L - Illumina 1.8+ Phred+33,  scores des séquences brutes compris entre 0 et 41
```

A partir d'Illumina 1.8, encodage sanger

Quality score encoded with a single ASCII character

Problème: différentes manières de coder la qualité de la séquence selon la technologie de séquençage illumina utilisée

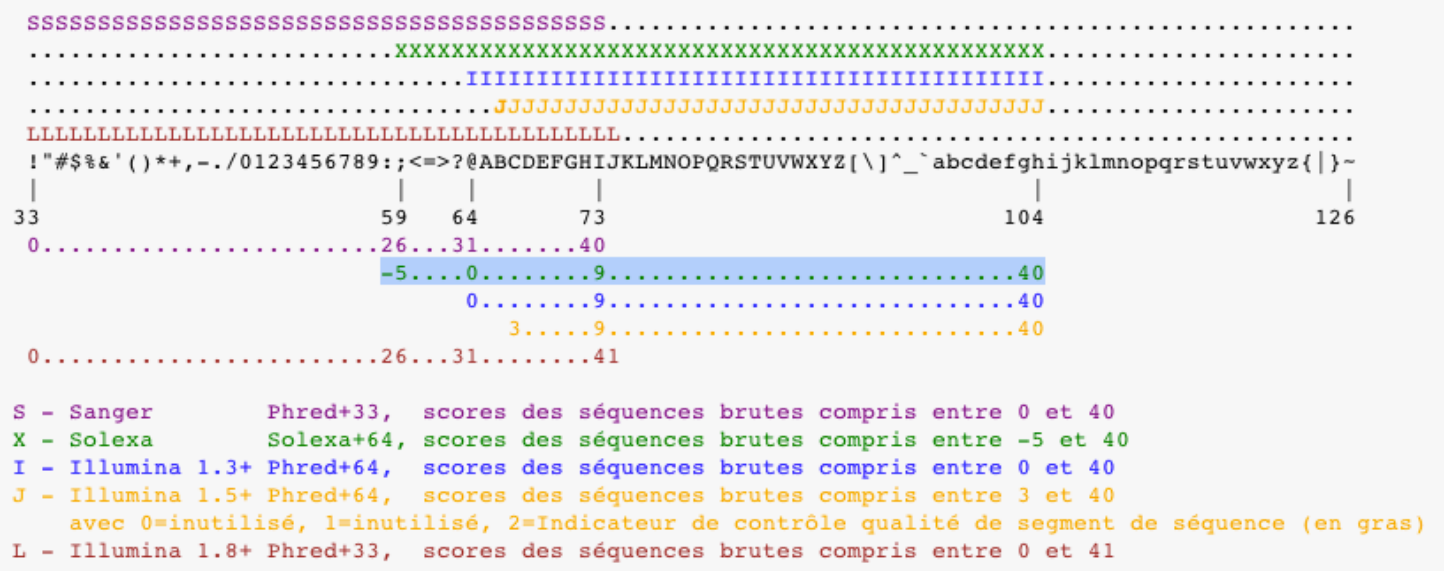
```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ..............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                               0........9.............................40
                                  3.....9.............................40
 0........................26...31.......41

S - Sanger        Phred+33,  scores des séquences brutes compris entre 0 et 40
X - Solexa        Solexa+64, scores des séquences brutes compris entre -5 et 40
I - Illumina 1.3+ Phred+64,  scores des séquences brutes compris entre 0 et 40
J - Illumina 1.5+ Phred+64,  scores des séquences brutes compris entre 3 et 40
    avec 0=inutilisé, 1=inutilisé, 2=Indicateur de contrôle qualité de segment de séquence (en gras)
L - Illumina 1.8+ Phred+33,  scores des séquences brutes compris entre 0 et 41
```

Be careful because the quality value depends on the quality encoding and the technology used

Quality score encoded with a single ASCII character

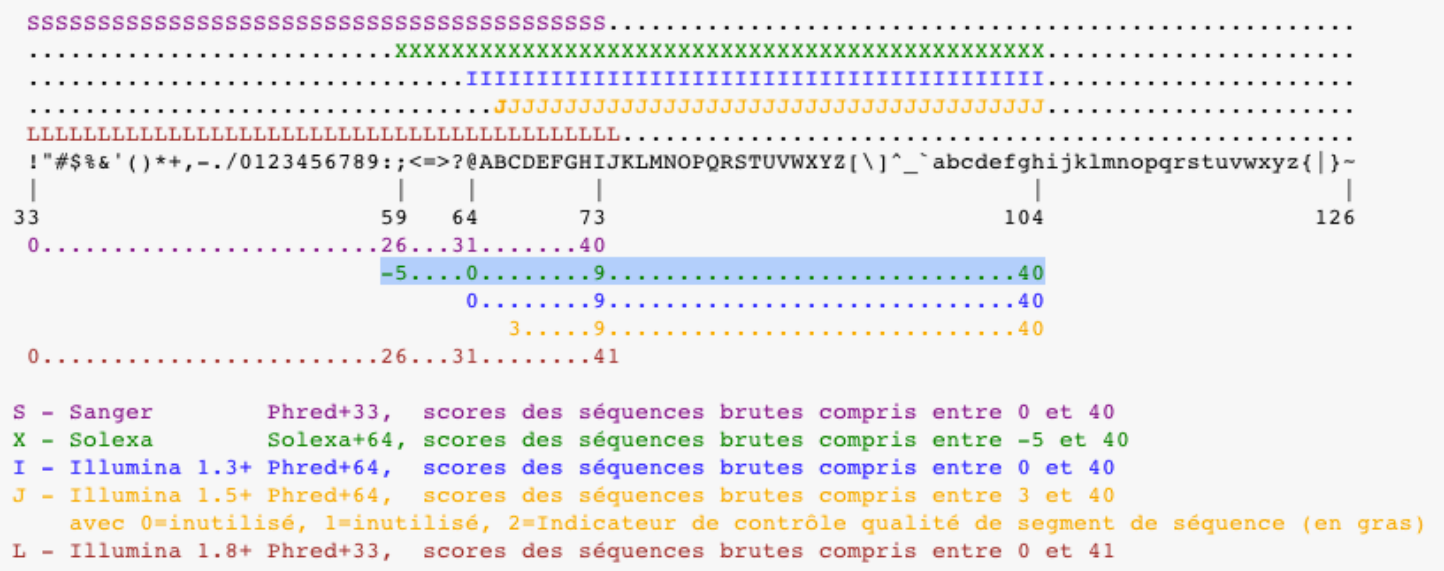
Problème: différentes manières de coder la qualité de la séquence selon la technologie de séquençage illumina utilisée



Fasqc donnera l'encodage utilisé lors du TP

Quality score encoded with a single ASCII character

Problème: différentes manières de coder la qualité de la séquence selon la technologie de séquençage illumina utilisée



Toujours utilisé l'encodage sanger !

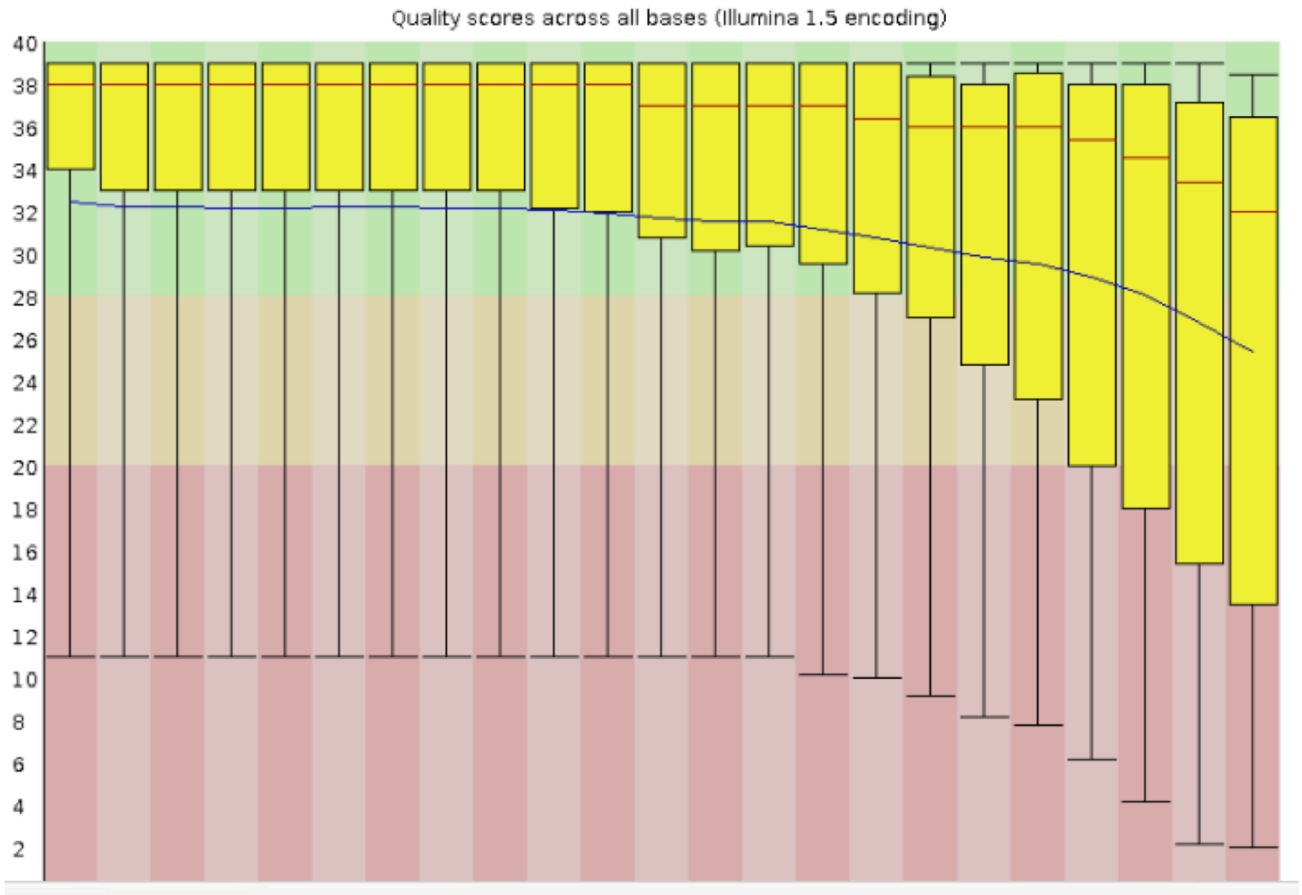
Quels logiciels utilisés?

- Contrôle de la qualité de séquençage

fastq -> ***fastqc***

Quality sequencing

Fastqc software



Quels logiciels utilisés?

- Contrôle de la qualité de séquençage

  fastq -> ***fastqc***

- “Nettoyage” des séquences :

  fastq -> ***cutadapt*** -> ***fastq***

- Mapping des séquences contre la référence sequence

  fastq -> ***bwa aln*** + ***bwa sampe*** -> ***sam*** -> ***bam***

SAM file format for Sequence Alignment Map

SAM format : http://samtools.sourceforge.net/samtools.shtml

Type	Tag	Description
HD - header	VN*	File format version.
	SO	Sort order. Valid values are: *unsorted*, *queryname* or *coordinate*.
	GO	Group order (full sorting is not imposed in a group). Valid values are: *none*, *query* or *reference*.
SQ - Sequence dictionary	SN*	Sequence name. Unique among all sequence records in the file. The value of this field is used in alignment records.
	LN*	Sequence length.
	AS	Genome assembly identifier. Refers to the reference genome assembly in an unambiguous form. Example: HG18.
	M5	MD5 checksum of the sequence in the uppercase (gaps and space are removed)
	UR	URI of the sequence
	SP	Species.
RG - read group	ID*	Unique read group identifier. The value of the ID field is used in the RG tags of alignment records.
	SM*	Sample (use pool name where a pool is being sequenced)
	LB	Library
	DS	Description
	PU	Platform unit (e.g. lane for Illumina or slide for SOLiD); should be a full, unambiguous identifier
	PI	Predicted median insert size (maybe different from the actual median insert size)
	CN	Name of sequencing center producing the read.
	DT	Date the run was produced (ISO 8601 date or date/time).
	PL	Platform/technology used to produce the read.
PG - Program	ID*	Program name
	VN	Program version
	CL	Command line
CO - comment		One-line text comments

SAM file format for Sequence Alignment Map

SAM format : http://samtools.sourceforge.net/samtools.shtml

Type	Tag	Description
HD - header	VN*	File format version.
	SO	Sort order. Valid values are: *unsorted*, *queryname* or *coordinate*.
	GO	Group order (full sorting is not imposed in a group). Valid values are: *none*, *query* or *reference*.
SQ - Sequence dictionary	SN*	Sequence name. Unique among all sequence records in the file. The value of this field is used in alignment records.
	LN*	Sequence length.
	AS	Genome assembly identifier. Refers to the reference genome assembly in an unambiguous form. Example: HG18.
	M5	MD5 checksum of the sequence in the uppercase (gaps and space are removed)
	UR	URI of the sequence
	SP	Species.
RG - read group	ID*	Unique read group identifier. The value of the ID field is used in the RG tags of alignment records.
	SM*	Sample (use pool name where a pool is being sequenced)
	LB	Library
	DS	Description
	PU	Platform unit (e.g. lane for Illumina or slide for SOLiD); should be a full, unambiguous identifier
	PI	Predicted me
	CN	Name of seq
	DT	Date the run
	PL	Platform/tech
PG - Program	ID*	Program nar
	VN	Program ver
	CL	Command lir
CO - comment		One-line text

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA    *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC     *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

SAM file format for Sequence Alignment Map

SAM format : http://samtools.sourceforge.net/samtools.shtml

Col	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	extended CIGAR string (operations: MIDNSHP)
7	NRNM	Mate Reference NaMe (`=' if same as RNAME)
8	MPOS	1-based leftmost Mate POSition
9	ISIZE	inferred Insert SIZE
10	SEQ	query SEQuence on the reference
11	QUAL	query QUALity (ASCII-33

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA    *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC     *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

SAM file format for Sequence Alignment Map

SAM format : http://samtools.sourceforge.net/samtools.shtml

Suppose we have the following alignment with bases in lower cases clipped from the alignment. Read r001/1 and r001/2 constitute a read pair; r003 is a chimeric read; r004 represents a split alignment.

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                           ttagctTAGGC
-r001/2                                      CAGCGCCAT
```

SAM file format for Sequence Alignment Map

SAM format : http://samtools.sourceforge.net/samtools.shtml

Suppose we have the following alignment with bases in lower cases clipped from the alignment. Read r001/1 and r001/2 constitute a read pair; r003 is a chimeric read; r004 represents a split alignment.

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                      ATAGCT..............TCAGC
-r003                            ttagctTAGGC
-r001/2                                         CAGCGCCAT
```

The corresponding SAM format is:

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       *  0   0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       *  0   0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         =  7 -39 CAGCGCCAT         *
```

Quels logiciels utilisés?

- Contrôle de la qualité de séquençage

  fastq -> ***fastqc***

- “Nettoyage” des séquences :

  fastq -> ***cutadapt -> fastq***

- Mapping des séquences contre la référence sequence

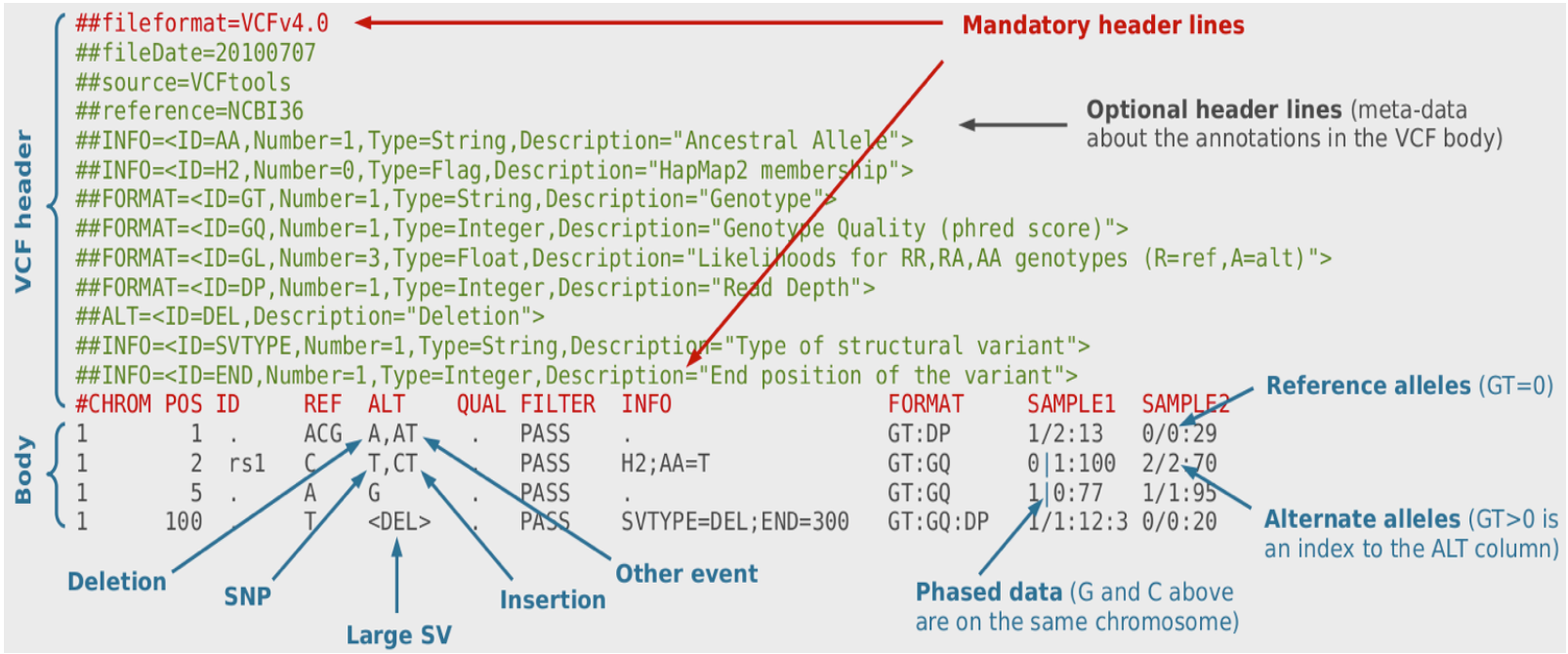
  fastq -> ***bwa aln*** + ***bwa sampe*** -> ***sam*** -> ***bam***

- Détection des SNP

  bam -> ***samtools mpileup***-> ***bcf*** -> ***bcftools*** -> ***vcf***

VCF file format for Variant Call Format

The Variant Call Format (VCF) used in bioinformatics for storing gene sequence variations



VCF file format for Variant Call Format

```
##fileformat=VCFv4.1
##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                              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''
```

- **Variation 1 :** a good SNP
- **Variation 2 :** a possible SNP that has been filtered out because its quality is below 10
- **Variation 3 :** a site at which two alternate alleles are called, with one of them (T) being ancestral (possibly a reference sequencing error)
- **Variation 4 :** a site that is called monomorphic reference (i.e. with no alternate alleles)
- **Variation 5 :** a microsatellite with two alternative alleles, one a deletion of 2 bases (TC), and the other an insertion of one base (T).