

The logo of the University of Bordeaux is displayed against a background with a blue diagonal stripe in the top left and a dark grey diagonal stripe in the bottom right. The text 'université' is in a dark brown, lowercase, sans-serif font, with a blue accent on the 'u' and 'e'. Below it, 'de' is in a smaller, dark brown, lowercase, sans-serif font, and 'BORDEAUX' is in a larger, dark brown, uppercase, sans-serif font.

université
de **BORDEAUX**

Introduction au TD NGS pour la cancérologie



Plan général sur les 8h de TD

Session 1 (3h)

- Installation des packages R
- Introduction des bases en cancérologie
- Présentation du contexte du TD et introduction aux analyses NGS
- TD partie 1: Traitement primaire des données NGS

Session 2 (3h)

- TD partie 2: Recherche de variants et classification de patients

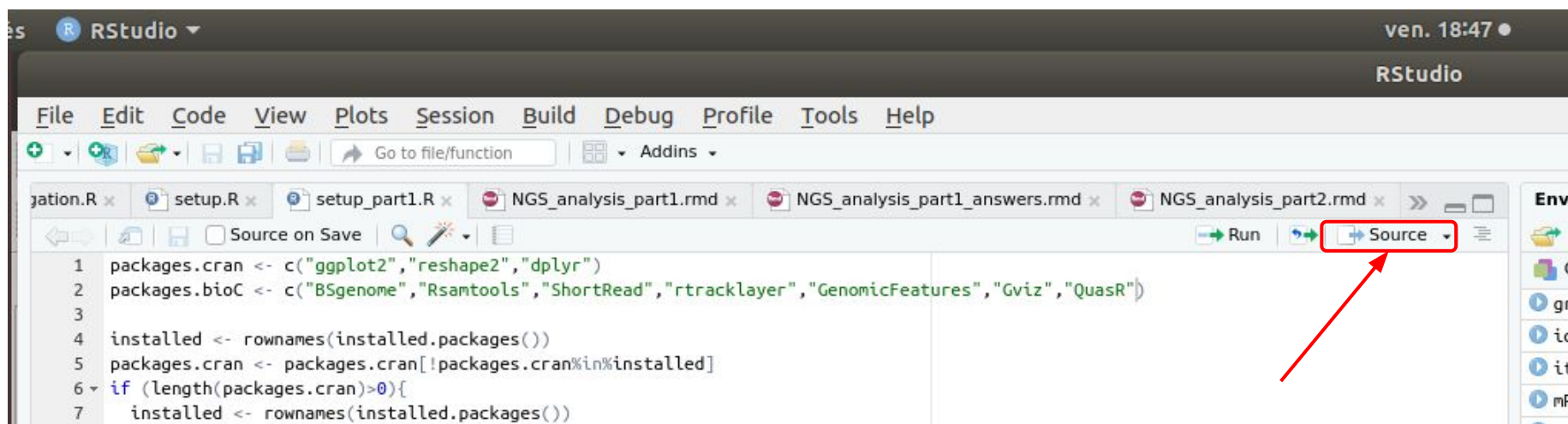
Session 3 (2h)

- TD partie 3: Analyse différentielle du transcriptome et annotations fonctionnelles

Installation des packages

Dans le dossier téléchargé vous trouverez le dossier *scripts*

- Ouvrez setup_part1.R dans RStudio
- Appuyez sur **Source**: ceci va lire tout le fichier et procéder aux installations



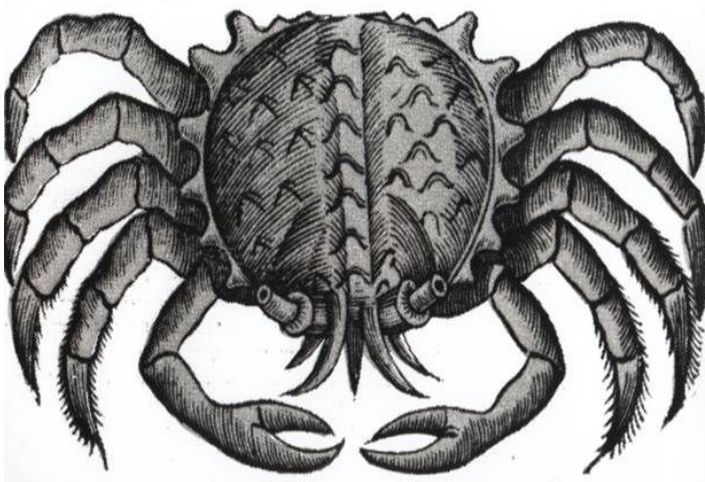
Pendant que ça installe ... Introduction générale

- Qu'est ce qu'un cancer ?
- Apport des Next Generation Sequencing (NGS)
- Pipeline de traitement des données
- Présentation du TD

Qu'est ce qu'un cancer ?

1. Histoire

- › 460 avant JC: Hippocrate nomme cette maladie "*karcinos*" (crabe) en analogie entre les ramifications des cellules cancéreuses et les pinces de crabe.



Qu'est ce qu'un cancer ?

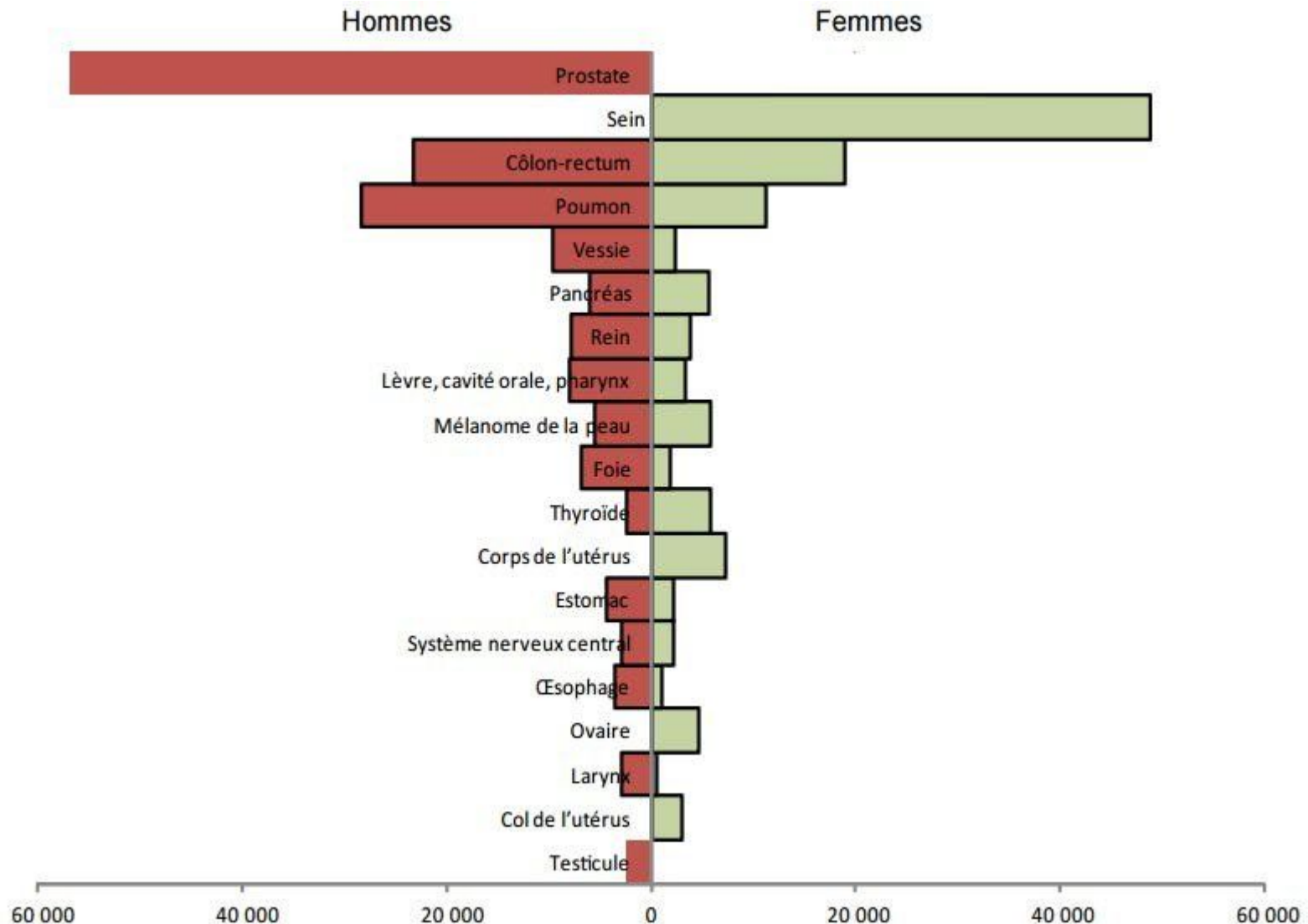
1. Histoire

- › 460 avant JC: Hippocrate nomme cette maladie "*karcinos*" (crabe) en analogie entre les ramifications des cellules cancéreuses et les pinces de crabe.

2. Incidence

- › En 2018 en France: ~382 000 nouveaux cas / ~157 000 décès (source: <https://www.santepubliquefrance.fr/maladies-et-traumatismes/cancers>)

Incidence des cancers en France (2012)



Muller, Marianna. (2017). Les ADN tumoraux circulants et leur rôle dans les cancers colorectaux métastatiques : application en théranostique.

Qu'est ce qu'un cancer ?

1. Histoire

- › 460 avant JC: Hippocrate nomme cette maladie "*karcinos*" (crabe) en analogie entre les ramifications des cellules cancéreuses et les pinces de crabe.

2. Incidence

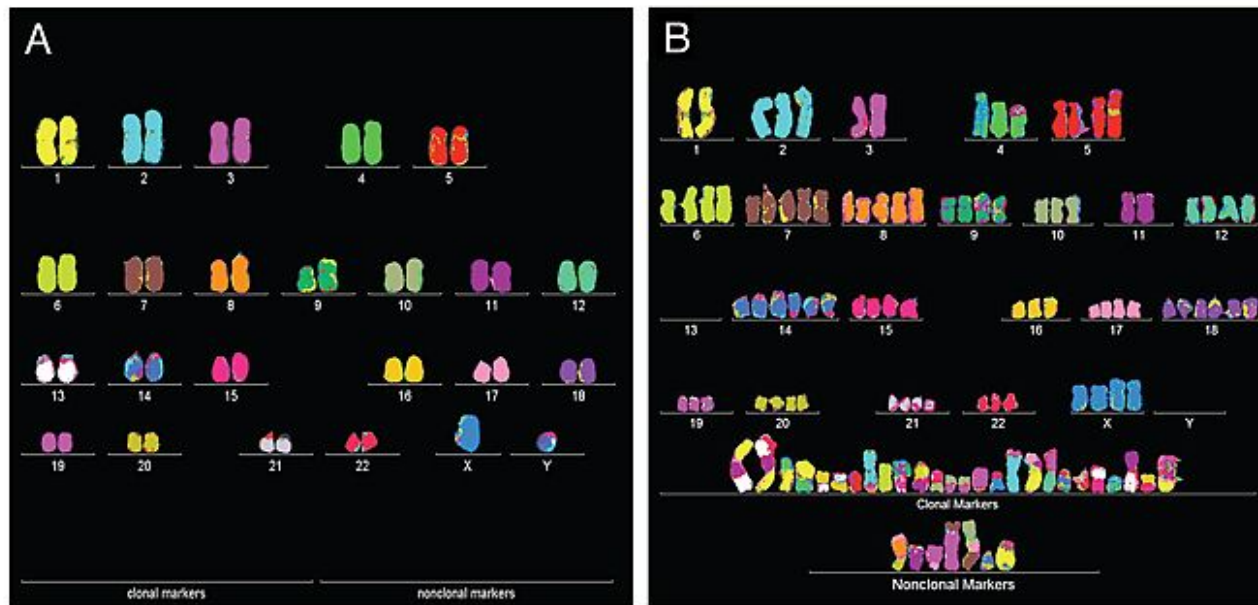
- › En 2018 en France: ~382 000 nouveaux cas / ~157 000 décès (source: <https://www.santepubliquefrance.fr/maladies-et-traumatismes/cancers>)

3. Définition

- › Amas de cellules immortelles (tumeur) se divisant anarchiquement qui altère le fonctionnement de l'organe où il se situe et peut se propager via le système sanguin ou lymphatique → métastases

Qu'est ce qu'un cancer ?

Un caryotype anormal: cellule normale (à gauche) et cancéreuse (cancer de la vessie) à droite, obtenus avec la technique de FISH



<https://www.berkeley.edu/>

De nombreuses anomalies apparaissent dans les cellules cancéreuses: nombre de chromosomes en plus ou en moins, translocations, délétions...La cellule cancéreuse a accumulé des mutations et des anomalies de la répartition des chromosomes

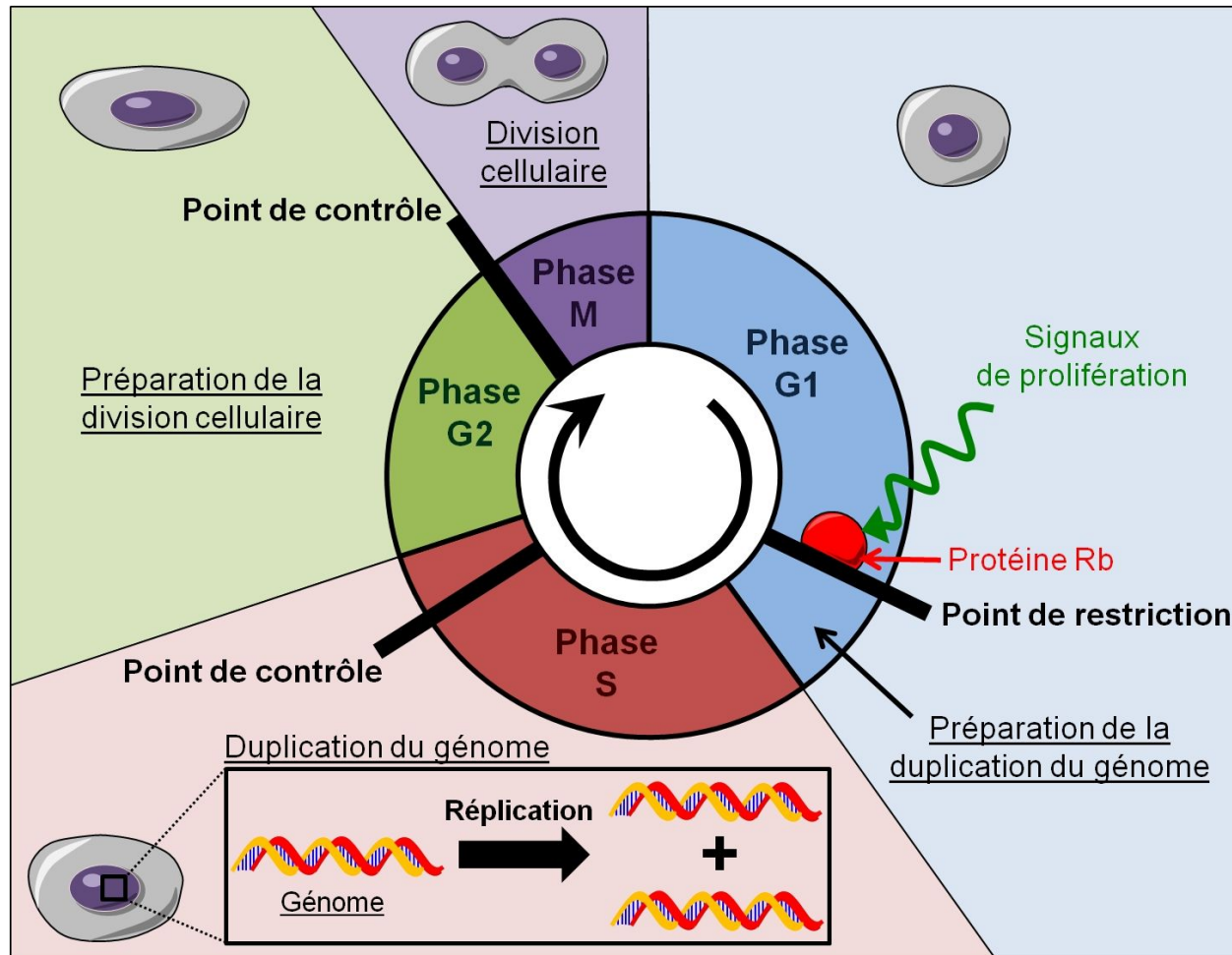
Les mutations acquises ou somatiques

- mutations ponctuelles silencieuses (polymorphismes): sans conséquence dans la plupart des cas.
- mutations ponctuelles délétères: conséquences sur l'expression des gènes, la conformation des protéines (troncage, changement d'acide aminé)
- variations structurales: délétions / duplications de gènes / régions entiers ou translocations (échanges) de chromosome ou de parties de chromosome qui peut mener à des fusion de gènes.

Acquisition des mutations

- › Risques endogènes: à chaque division cellulaire l'ADN est répliqué, des erreurs peuvent-être introduites
 - ADN: 3×10^9 nucléotides
 - taux de mutation: $\sim 10^{-8} \rightarrow \sim 100$ mutations / cycle cellulaire
 - > 1 milliard divisions cellulaires / jour (remplacement de cellules vieillissantes ou endommagées)
- › Risques exogènes: tabac, alcool, soleil, polluants, virus etc ...
- › Heureusement les cellules ont des mécanismes de réparation ou d'induction de leur mort (apoptose) contrôlés par des gènes.

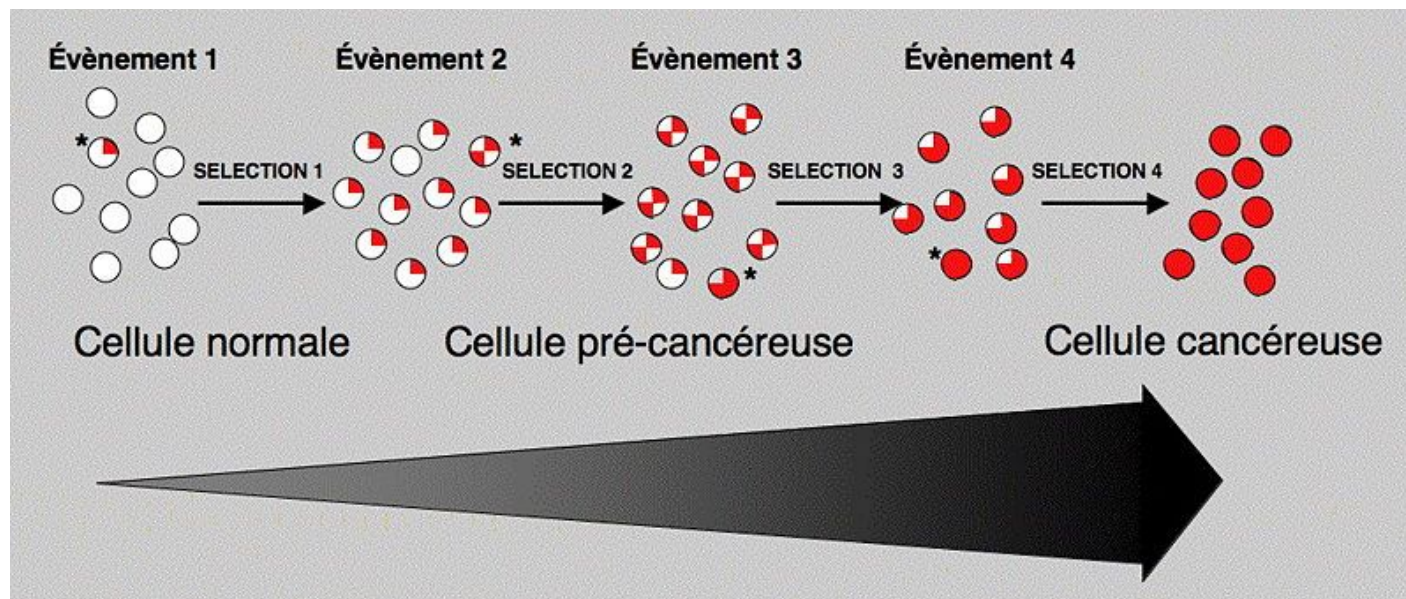
Le cycle cellulaire



<http://gregynours.e-monsite.com/pages/comment-apparait-une-cellule-cancereuse.html>

Transformation des cellules

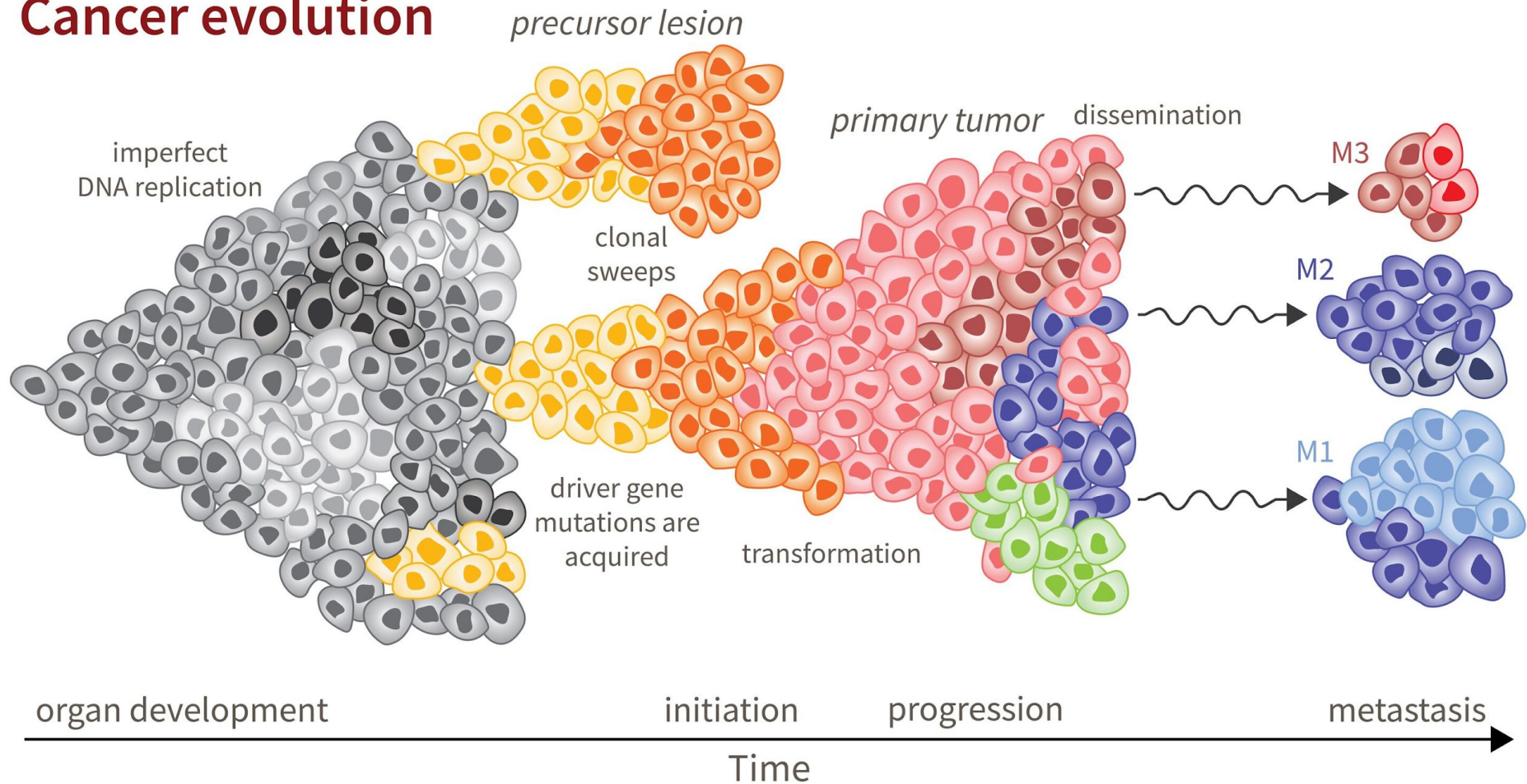
Dans de rares cas les mutations peuvent atteindre un gène réparateur ou contrôlant la multiplication cellulaire



http://fr.wikipedia.org/wiki/Fichier:Multi_Onco.jpg

Processus long qui peut prendre plusieurs années

Cancer evolution



Reiter Lab

Les gènes impliqués dans la cancérogenèse

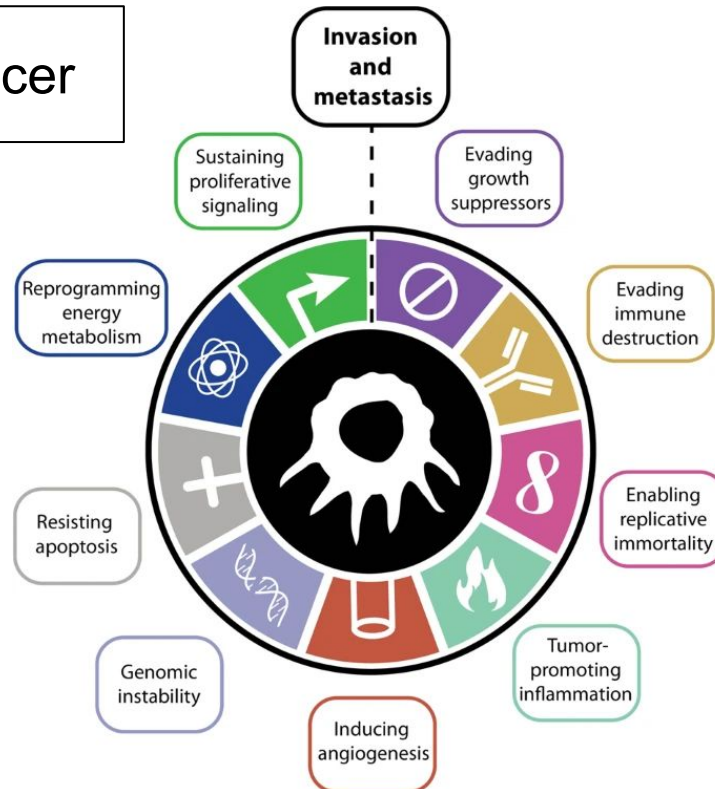
La cellule reçoit en permanence des signaux chimiques des autres cellules qui peuvent lui ordonner de se diviser, de se reposer ou de s'autodétruire.

- **proto-oncogènes**: normaux et indispensables, transmettent les signaux mitotiques → mutation → **oncogènes** → stimulation incontrôlée des divisions cellulaires
- **suppresseurs de tumeur**: freinent la prolifération des cellules (i.e. gène P53 muté dans >60% des cancers) → mutation → inactivation ou une diminution de leur fonctionnement → stimulation anormale de la prolifération cellulaire
- **réparateurs de l'ADN** → mutation → favorisent l'apparition des cancers.

Caractéristiques d'une cellule cancéreuse

- son indépendance vis-à-vis des signaux qui régulent (favorisent ou freinent) habituellement sa croissance et sa division
- sa capacité à échapper au processus de mort cellulaire programmée
- sa capacité à se diviser indéfiniment

The hallmarks of cancer



Challenge: détecter les mutations promotrices (marqueurs moléculaires)

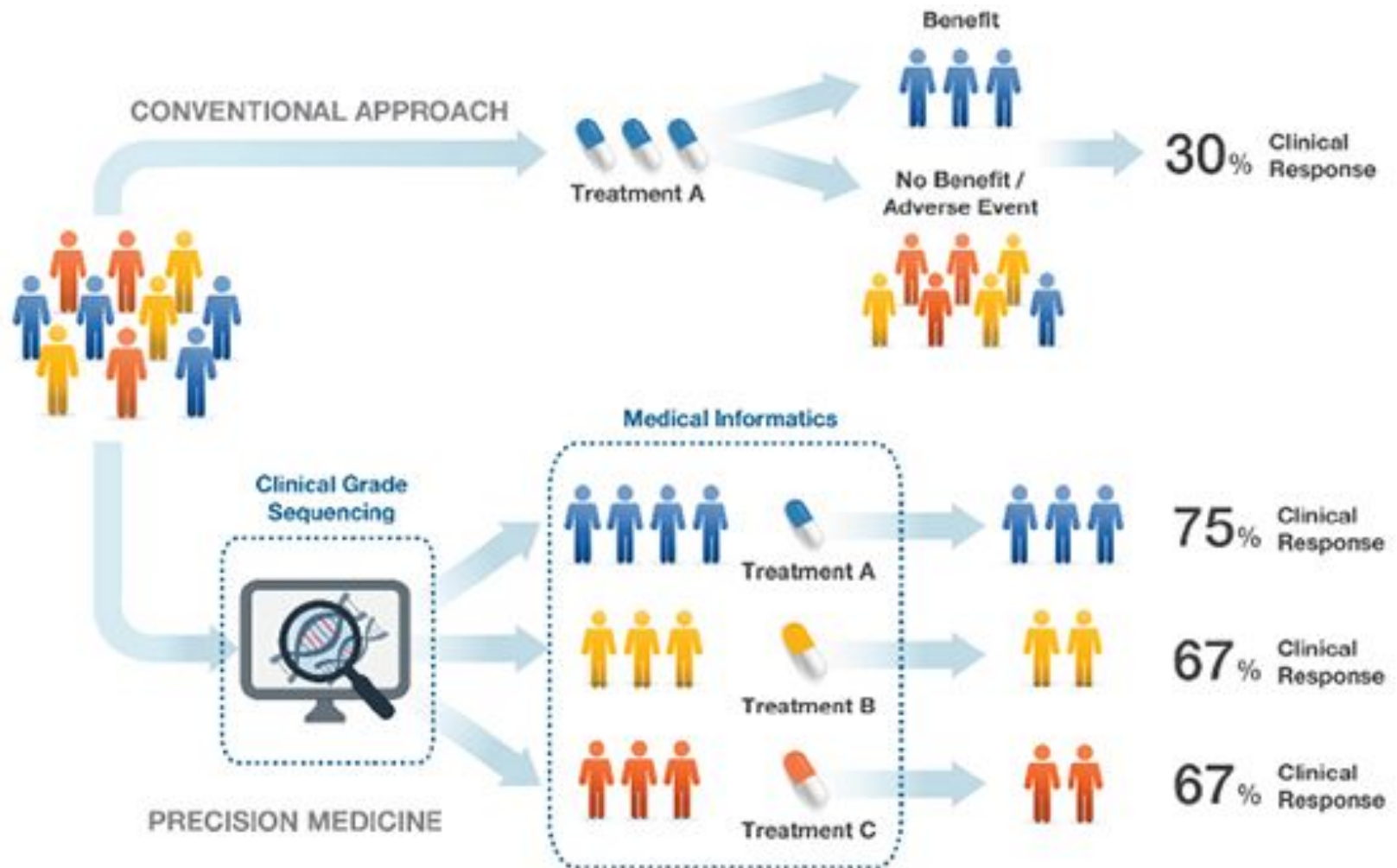
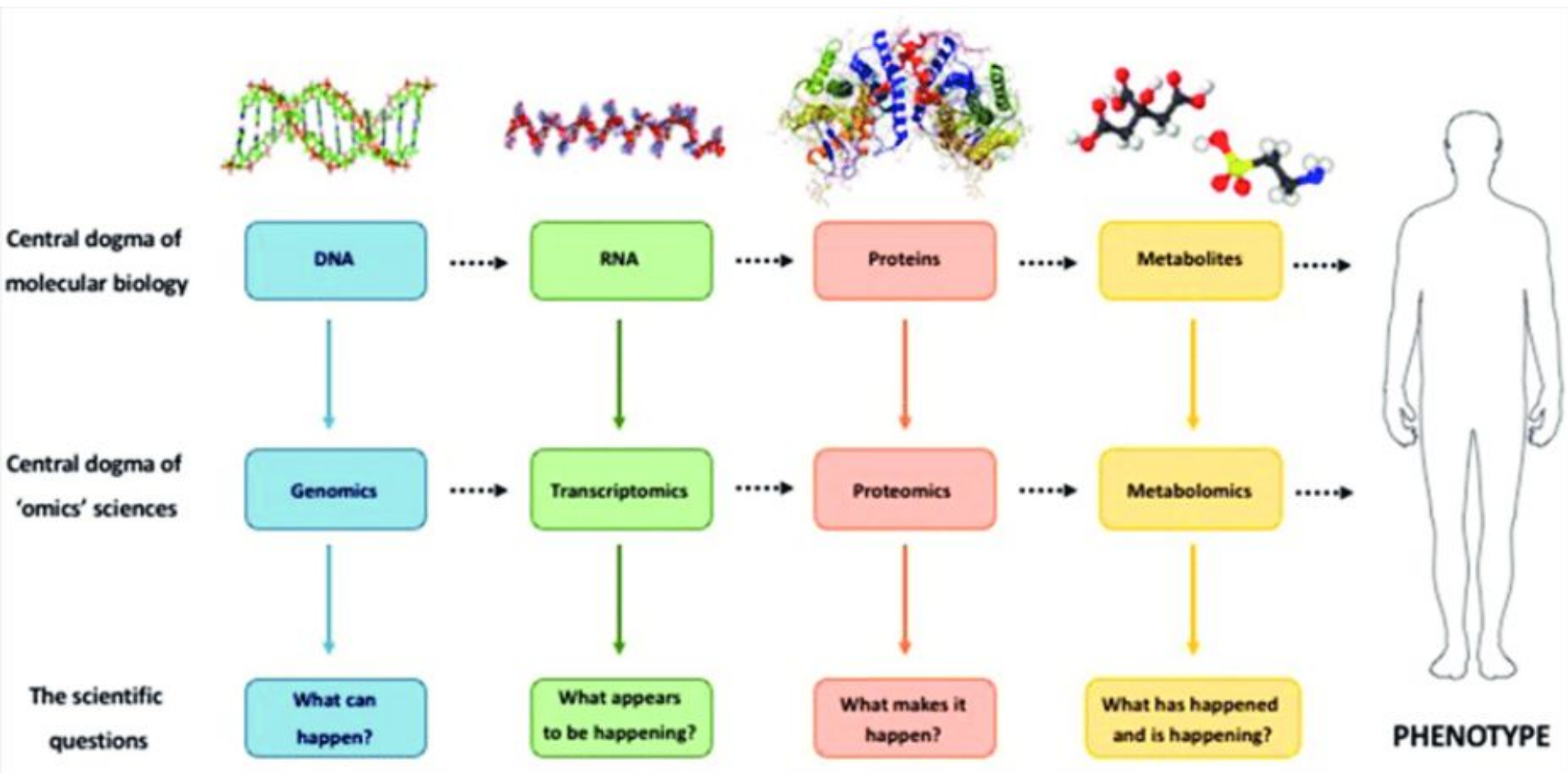


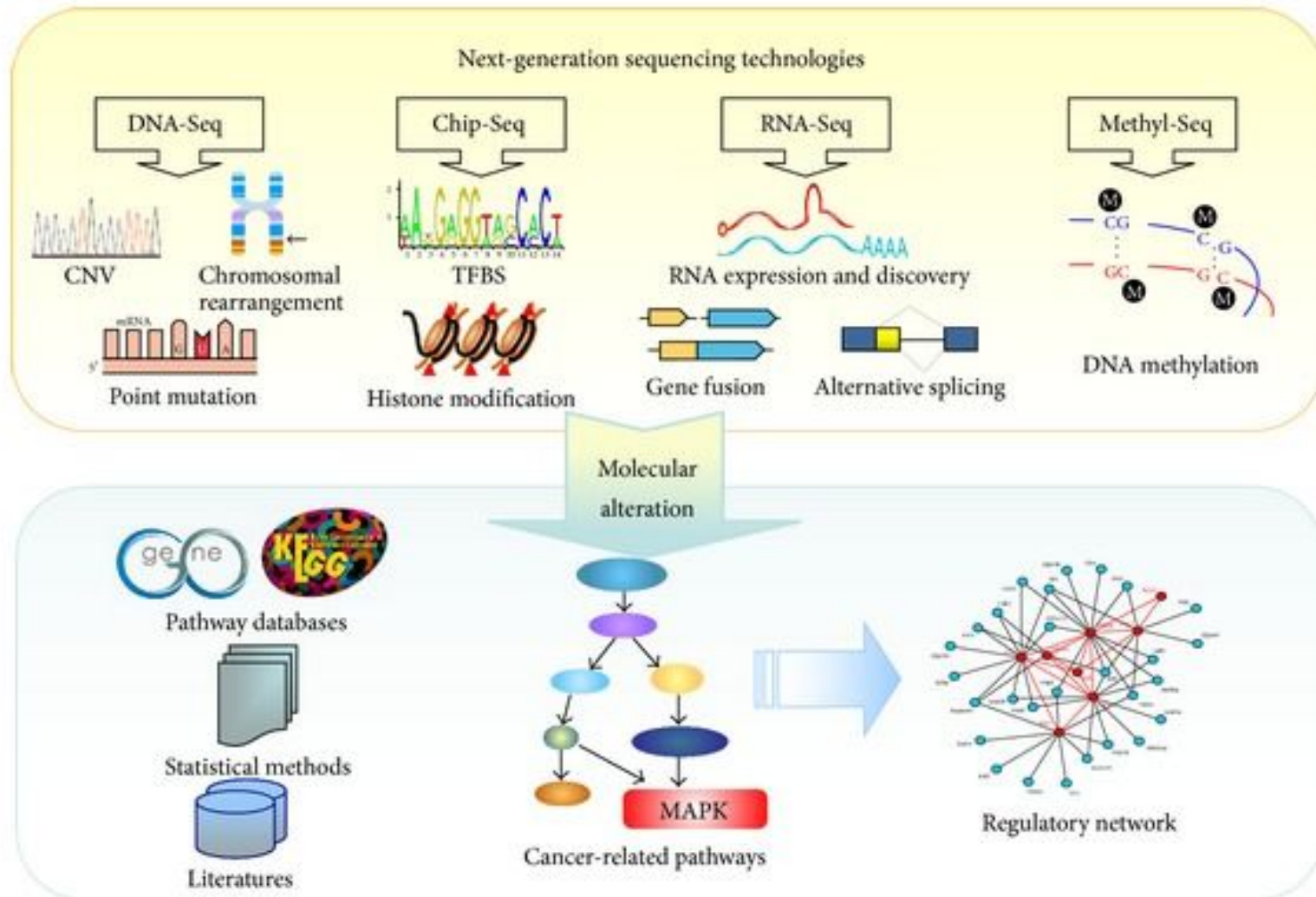
Figure 1. Precision diagnostics stratifies patients according to their molecular signature

L'utilité des données Next Generation Sequencing (NGS)

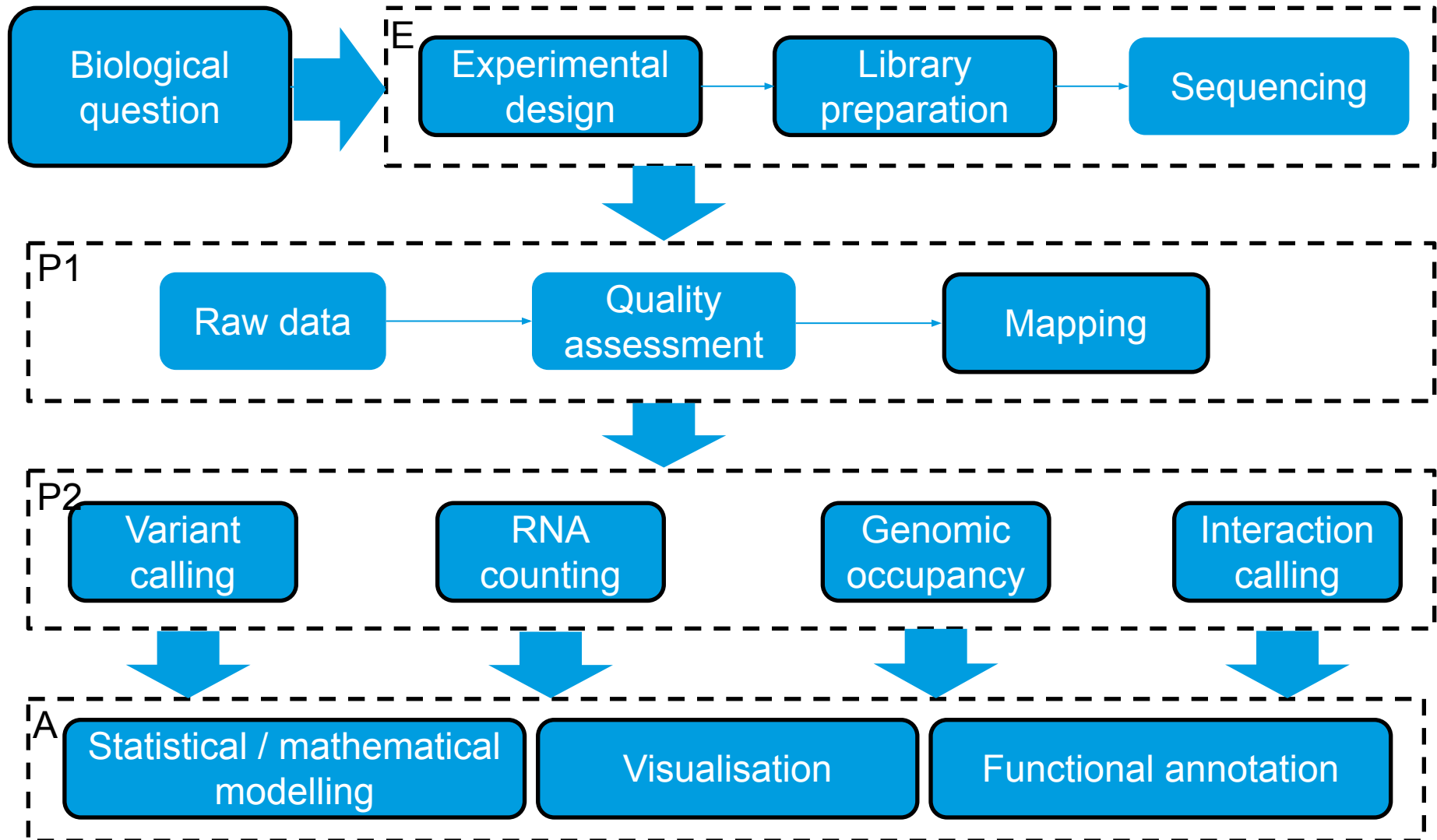


Araújo et al. 2017 Critical Reviews in Toxicology 47(8)

L'utilité des données Next Generation Sequencing (NGS)



Vue d'ensemble d'un projet



Contexte biologique:

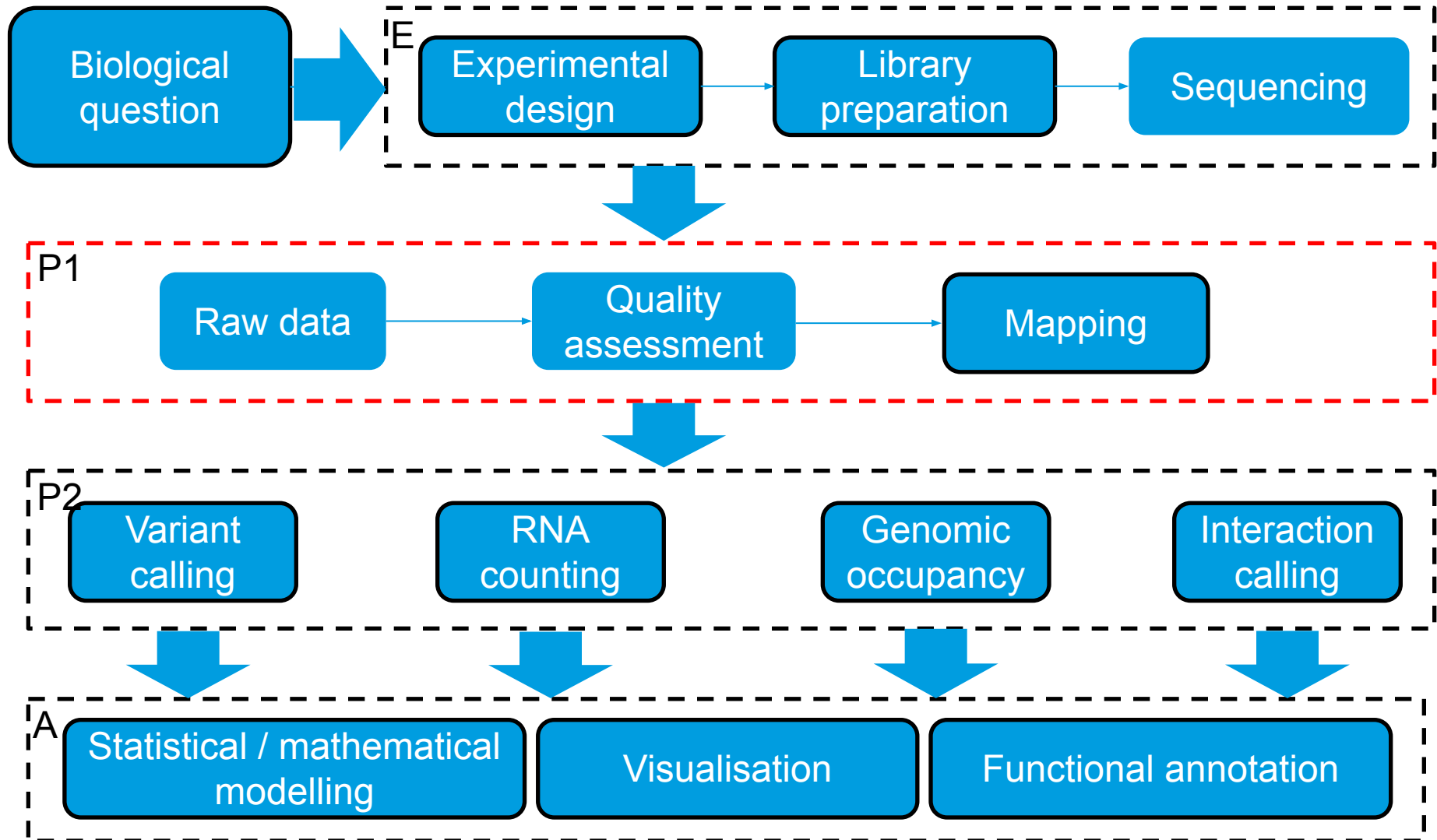
- Les léiomyosarcomes (LMS) sont des cancers des tissus mous
 - montrant une différenciation musculaire lisse
 - se trouvant à de multiples endroits du corps
- Aucun marqueur génétique ne permet de les identifier systématiquement
- Il n'existe pas de thérapie ciblée
- Génétique complexe inexpliquée
 - Seules les voies de signalisation Rb1 et TP53 ont été identifiées
 - N'explique pas tous les phénotypes tumoraux

Ce qu'on va faire sur l'ensemble du TD

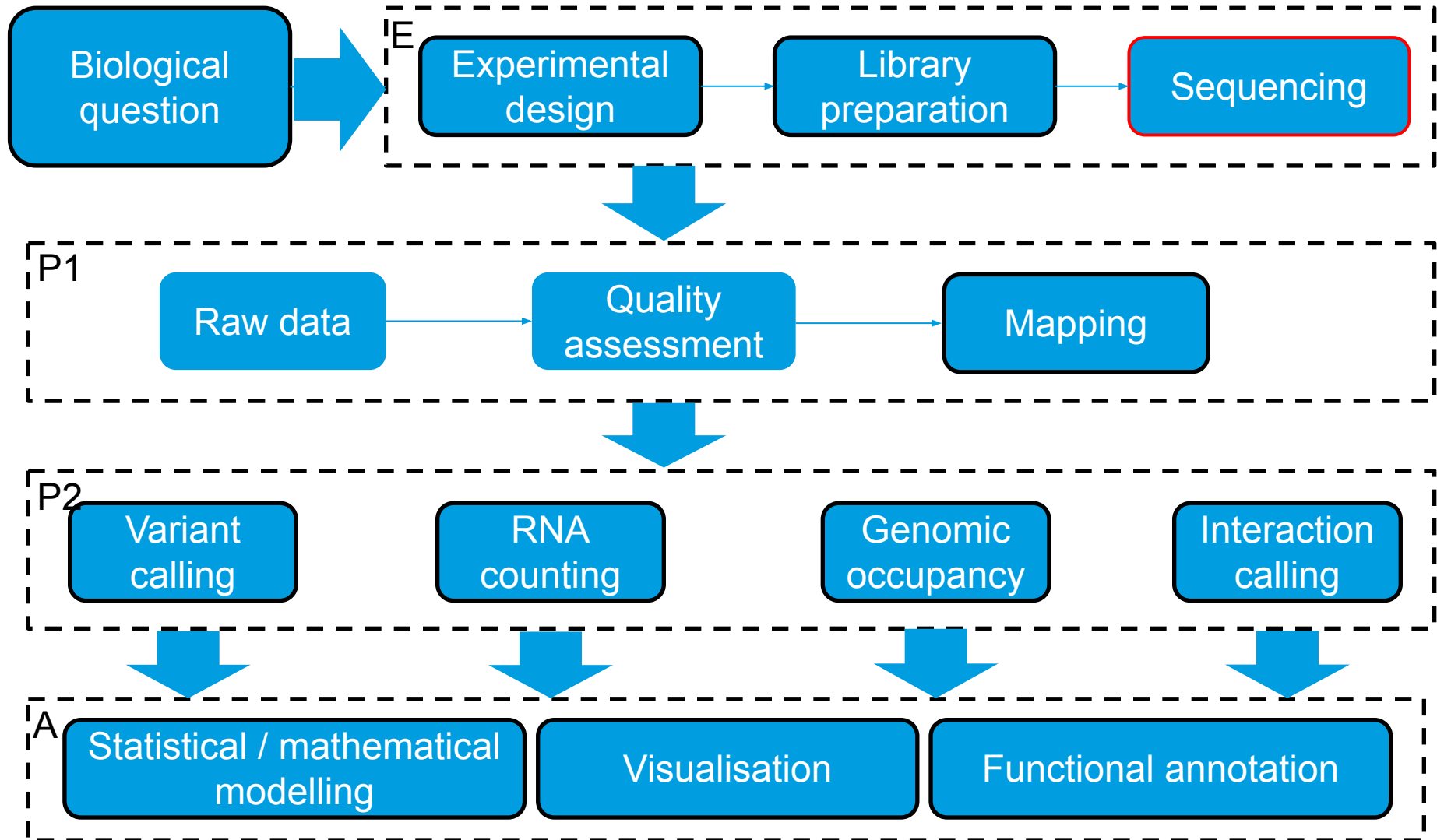
- Manipuler des fichiers de données NGS de différents types:
 - › Fastq
 - › Bam
 - › Vcf
- Utiliser RStudio
 - › Utilisation de commandes basiques et avancées
 - › Suivre un pipeline d'analysis
- Utiliser les NGS pour comprendre un phénotype
 - › Recherche de variants ponctuels
 - › Recherche de variants structuraux
 - › Classification des patients
 - › Recherche de gènes différentiellement exprimés et de voies de signalisation dérégulées

- Comprendre les principes derrière une analyse NGS
 - › Que veut dire qualité ?
 - › Quelles sont les étapes importantes ?
 - › Des 100aines de paramètres qui peuvent changer les résultats
 - › NO MAGIC !
- Comprendre l'intérêt des NGS et de leur analyse pour la cancérologie

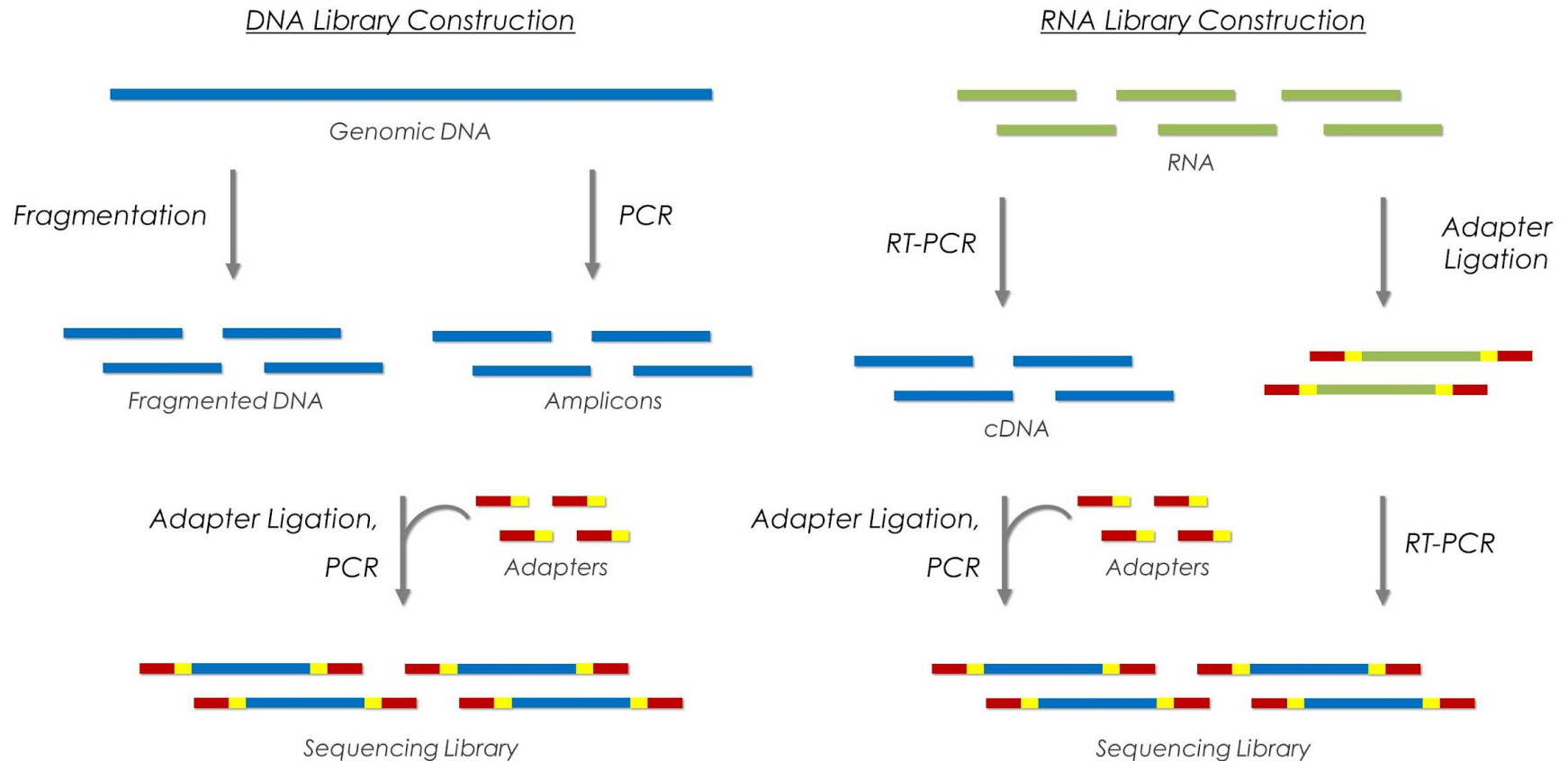
TD partie 1: étapes



TD partie 1: étapes

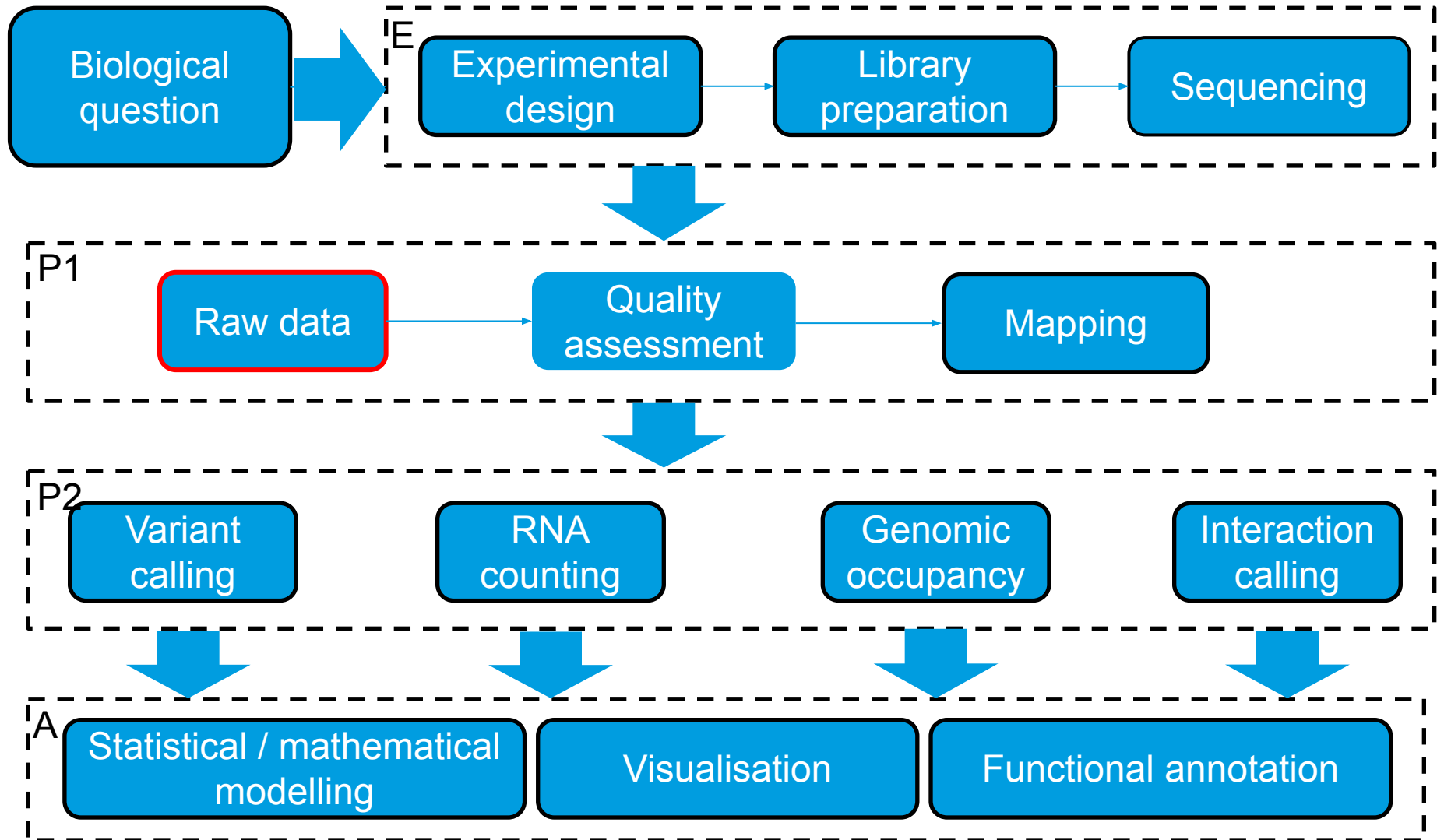


Production des données NGS



General overview of NGS Library Construction.

TD partie 1: étapes

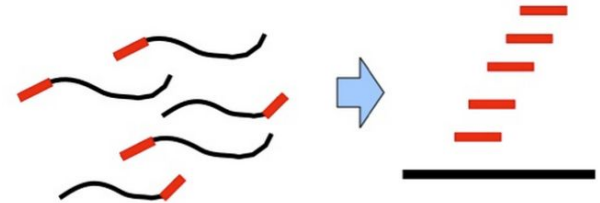


Données brutes: fastq files (.fastq ou .fq)

Fastq (.fastq ou .fq)

```
@SRR064166.142 HWI-EAS229_104:7:1:1:510 length=37/1  
GCAAAATGGATCCGTAACCTTCGGGAAAAGGATTGGCT  
+  
BB@?4@B@BAB@6@?B6AB@;)>*3/:2BB4B#####
```

Single-end (SE): from each cDNA fragment only one end is read

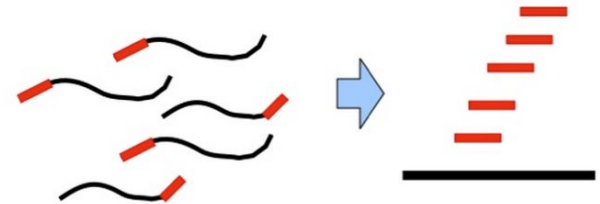


Données brutes: fastq files (.fastq ou .fq)

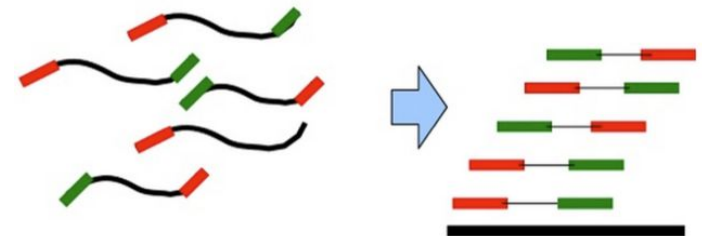
```
@SRR064166.142 HWI-EAS229_104:7:1:1:510 length=37/1
GCAAAATGGATCCGTAACCTTCGGGAAAAGGATTGGCT
+
BB@?4@B@BAB@6@?B6AB@;)>*3/:2BB4B#####
```

```
@SRR064166.152 HWI-EAS229_104:7:1:1:1001 length=37/1
GTTTAAGCATATCAATAAGCGGAGGAACAGACACTAA
+
BBCB>ACBABCCECA;BB5)>C5,.B6<1>C<1<6?CC
@SRR064166.152 HWI-EAS229_104:7:1:1:1001 length=37/2
CGACTTCCCTTGCCTACATTGTTCCATCGACCAGAGG
+
BBCBBCABBBBBBBBCA=BAA<AC=;@CA;?BB@5>##
```

Single-end (SE): from each cDNA fragment only one end is read



Paired-end (PE): the cDNA fragment is read from both ends



Données brutes: fastq files (.fastq ou .fq)

```
@SRR064166.142 HWI-EAS229_104:7:1:1:510 length=37/1
GCAAAATGGATCCGTAACCTTCGGGAAAAGGATTGGCT
+
BB@?4@B@BAB@6@?B6AB@;>)*3/:2BB4B#####
```

sequence identifier

sequence (IUPAC nomenclature)

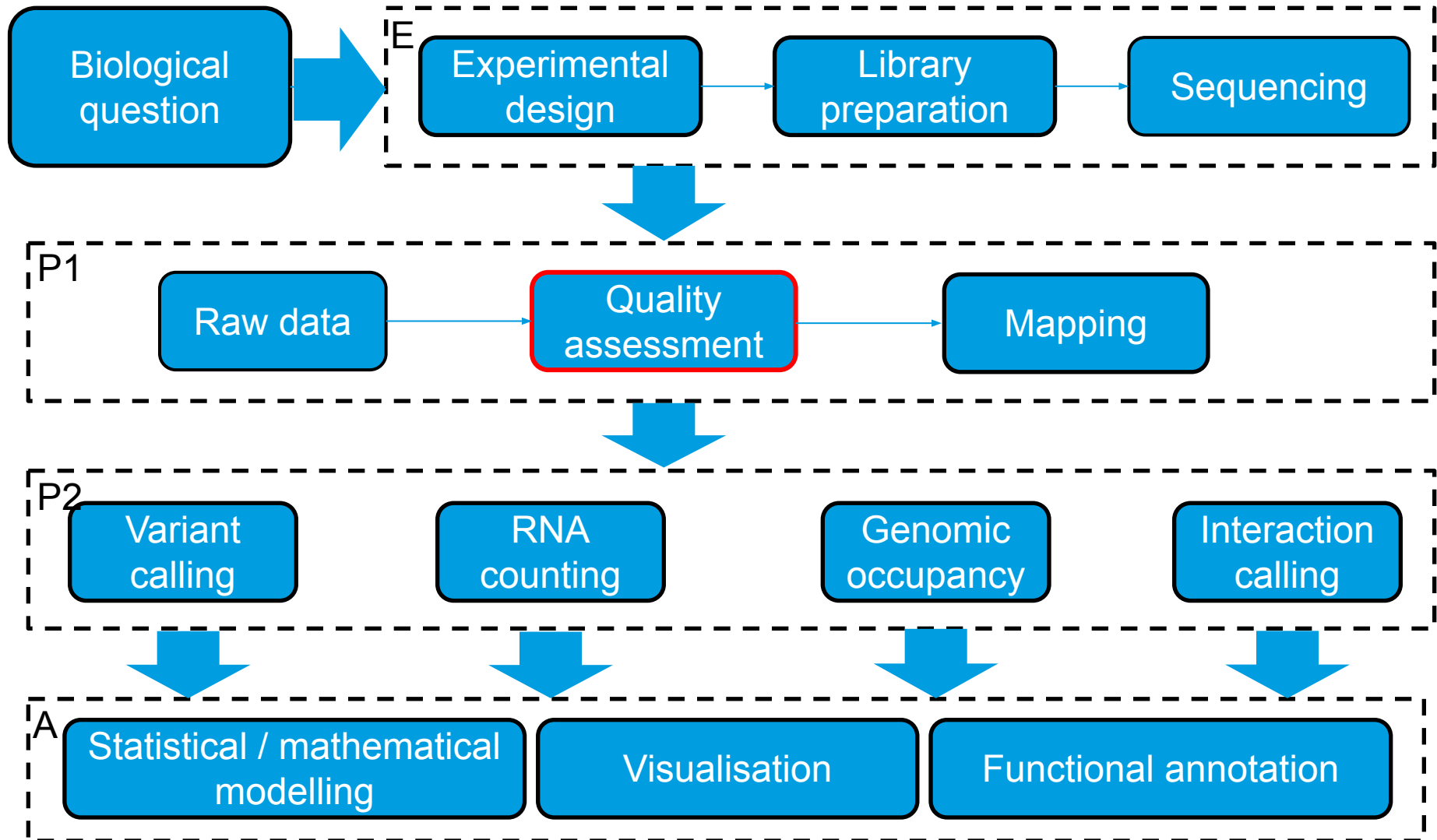
sequencing quality (ASCII encoding)

Correspond à un nombre qui peut être traduit en probabilité **p** que la base ait été **mal appelée**

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

| Phred Quality Score <i>Q</i> | Probability of incorrect base call <i>P</i> | Base call accuracy |
|------------------------------|---|--------------------|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1000 | 99.9% |
| 40 | 1 in 10,000 | 99.99% |
| 50 | 1 in 100,000 | 99.999% |
| 60 | 1 in 1,000,000 | 99.9999% |

TD partie 1: étapes



Qualité du séquençage: analyse d'un résultat fastQC

C'est la première étape et elle est **essentielle**.

Checklist:

- ❑ Format de fichier: statistiques basiques
- ❑ Fiabilité du base calling: Score qualité par base / par séquence
- ❑ Problème transitoire avec le run: Faible qualité dans les cellules
- ❑ Contamination: Contenu des séquences par base, séquences sur-représentées













Quelques outils:

- **FastQC** (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>)
- R/Bioconductor **ShortRead package** (fonctions qa() et report())
- R/Bioconductor **QuasR package** (fonction qQCReport())

File format

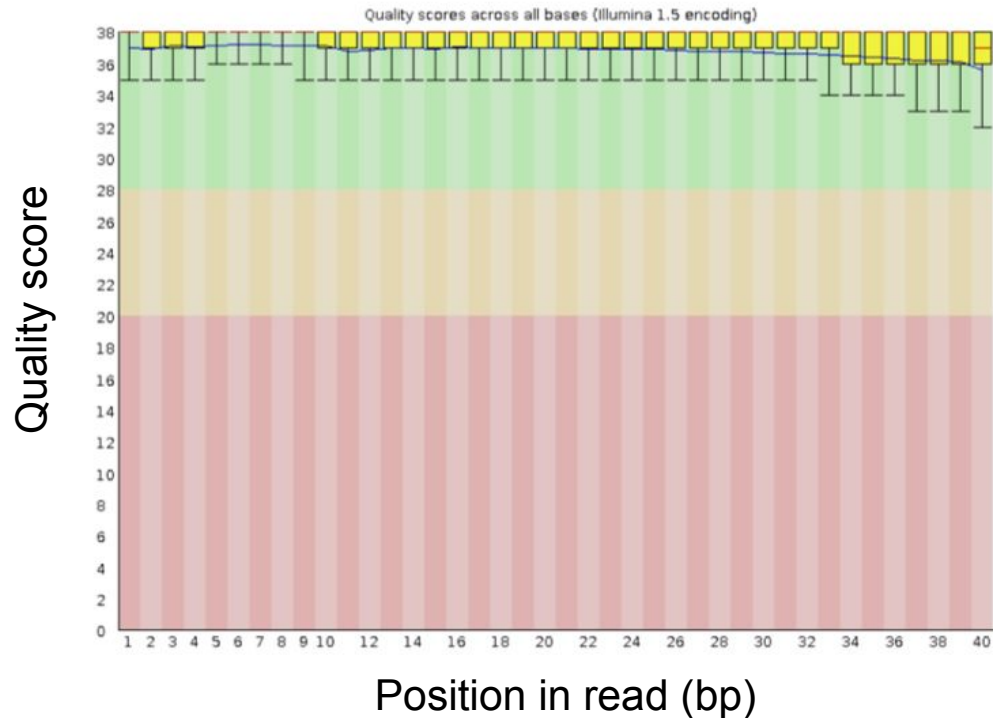
FastQC Report

Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per tile sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter Content](#)
-  [Kmer Content](#)

| Measure | Value |
|-----------------------------------|-------------------------|
| Filename | SRR064167.fastq |
| File type | Conventional base calls |
| Encoding | Sanger / Illumina 1.9 |
| Total Sequences | 17367742 |
| Sequences flagged as poor quality | 0 |
| Sequence length | 38 |
| %GC | 47 |

Reliability of base calling (per base)



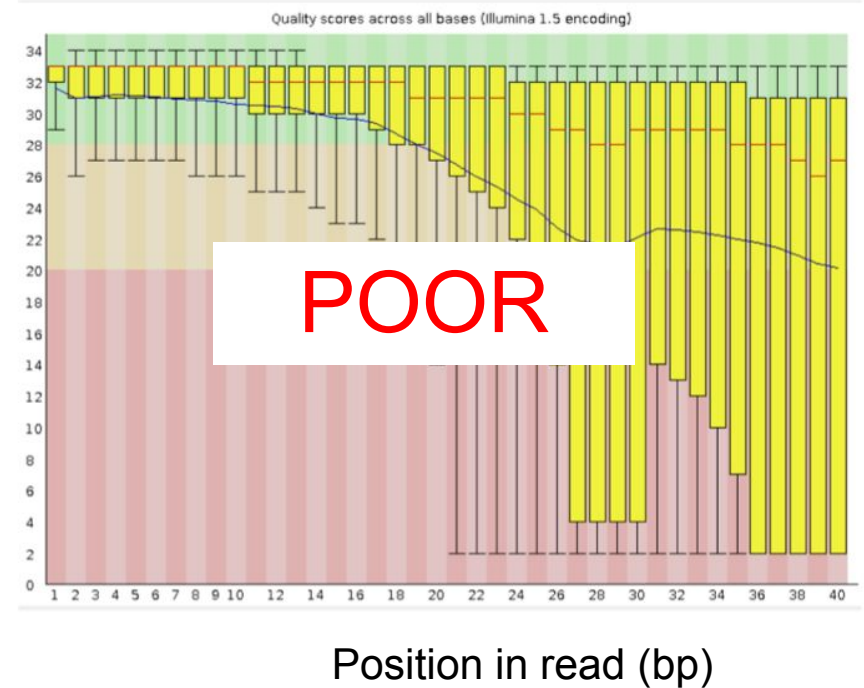
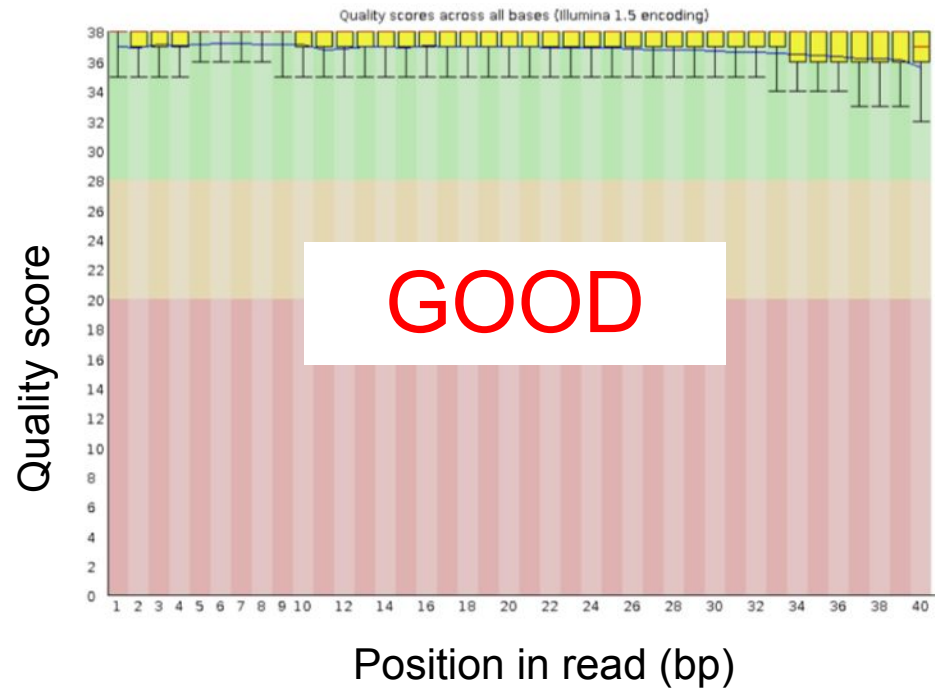
Quality score:
The **higher** the **better**

*Quality of calls from most platforms will **degrade** as the **run processes***

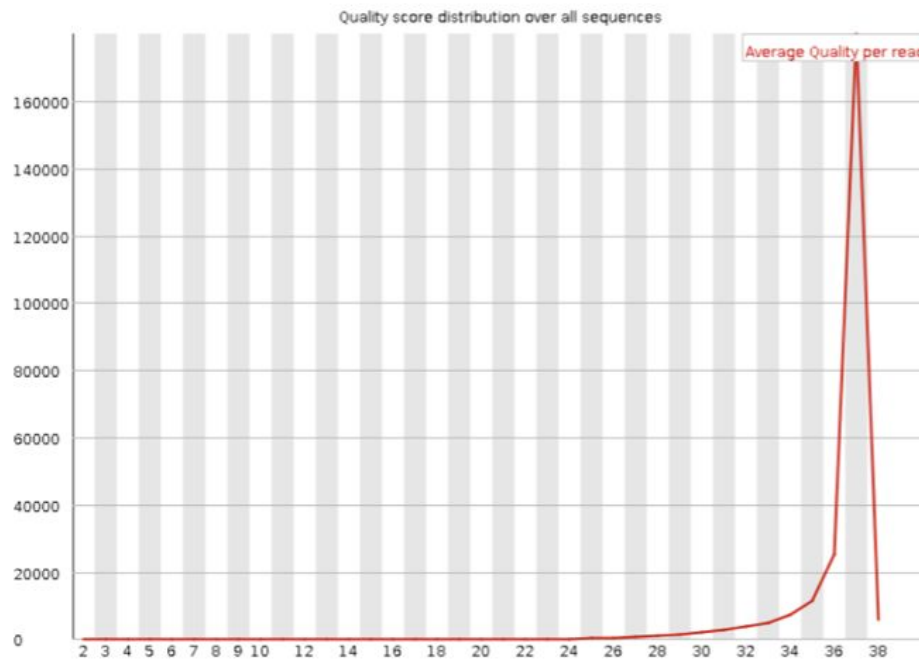
Warning: if the lower quartile for any base is less than 10, or if the median for any base is less than 25.

Failure: if the lower quartile for any base is less than 5 or if the median for any base is less than 20.

Reliability of base calling (per base)



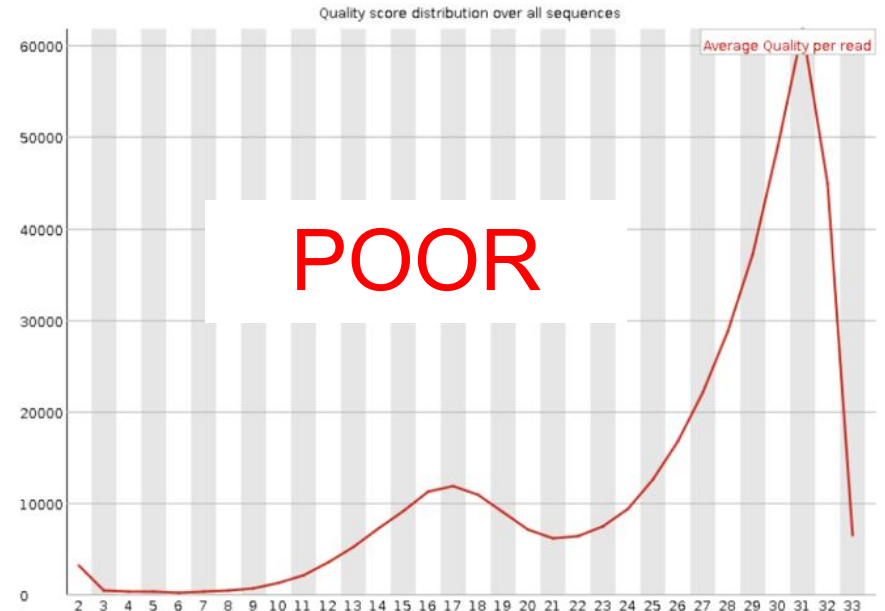
Reliability of base calling (per sequence)



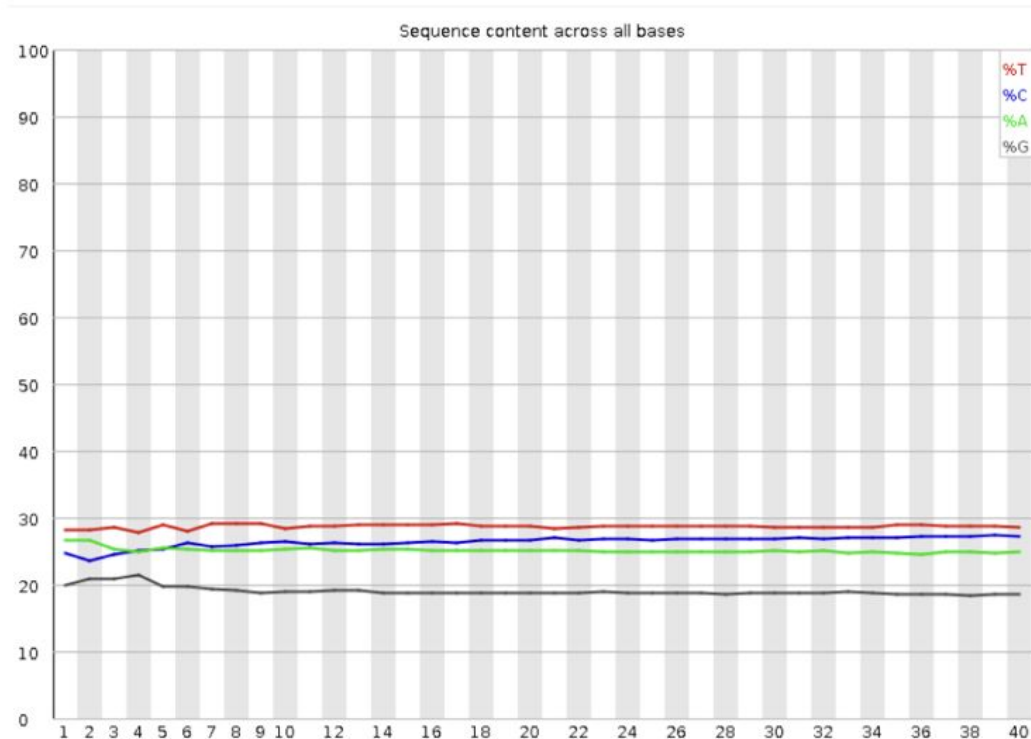
The per sequence quality score report allows you to see if a **subset** of your sequences have **universally low quality values**.

- Warning:** if the most frequently observed mean quality is below 27
- this equates to a 0.2% error rate.
- Failure:** if the most frequently observed mean quality is below 20
- this equates to a 1% error rate.

Reliability of base calling (per sequence)



Contamination: per base sequence content

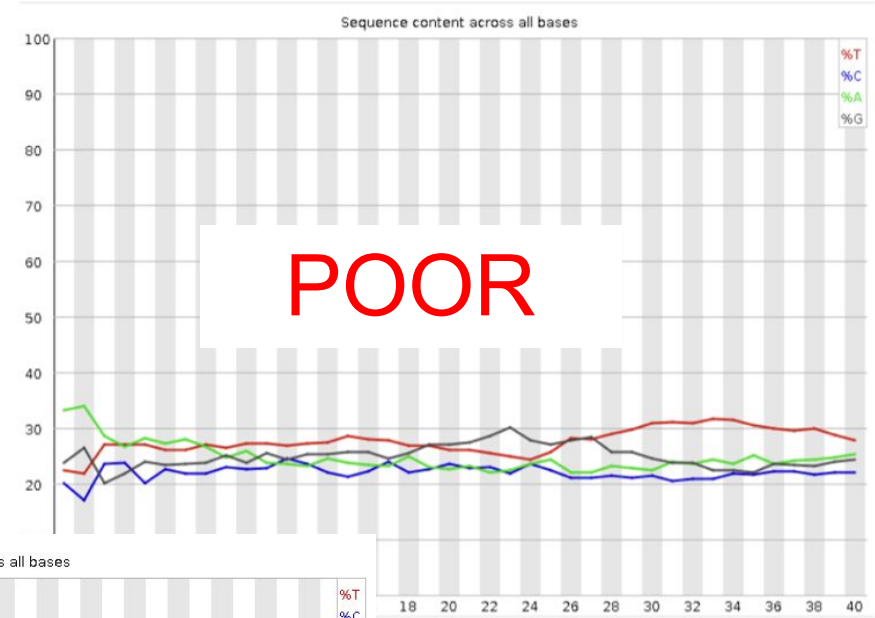
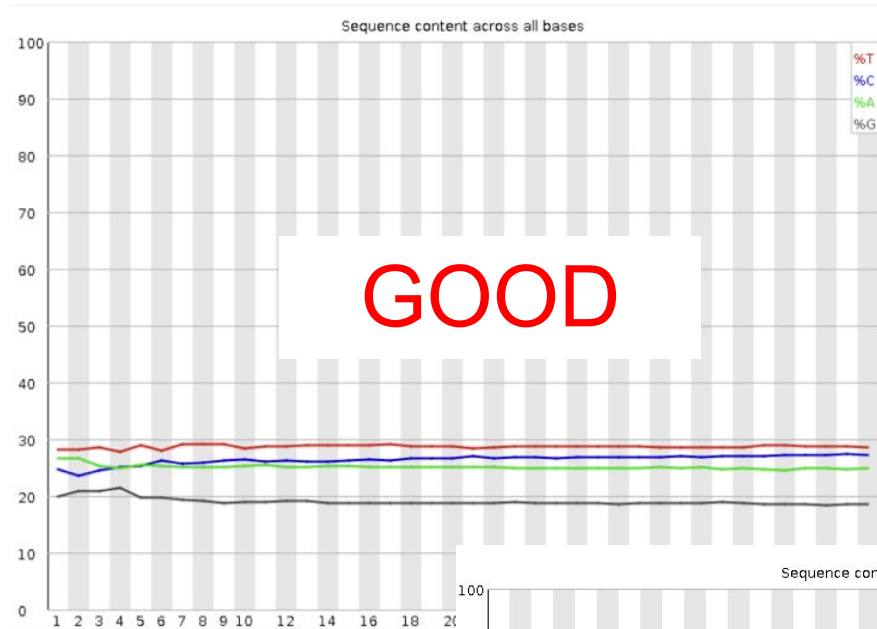


In a **random library** you would expect **no difference** between the different bases

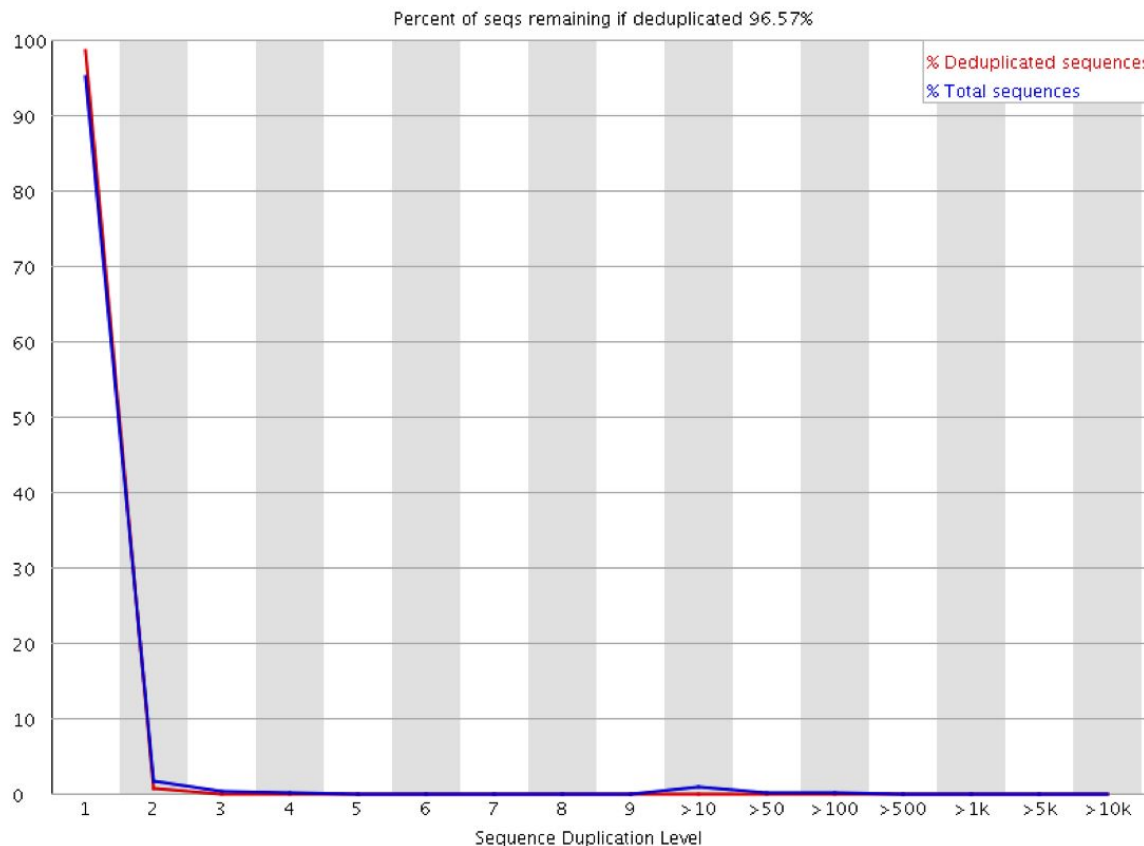
Warning: if the difference between A and T, or G and C is greater than 10% in any position.

Failure: if the difference between A and T, or G and C is greater than 20% in any position.

Contamination: per base sequence content



Duplication: subset enrichment



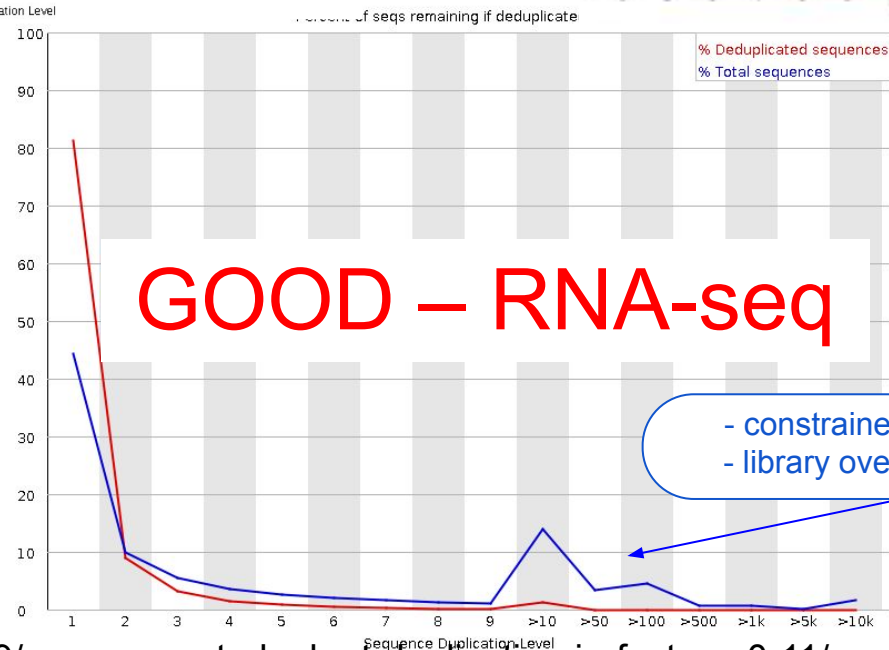
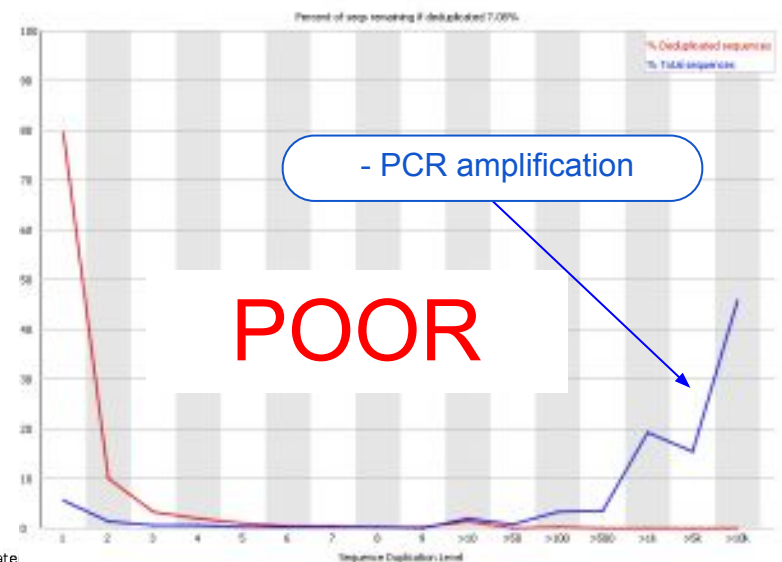
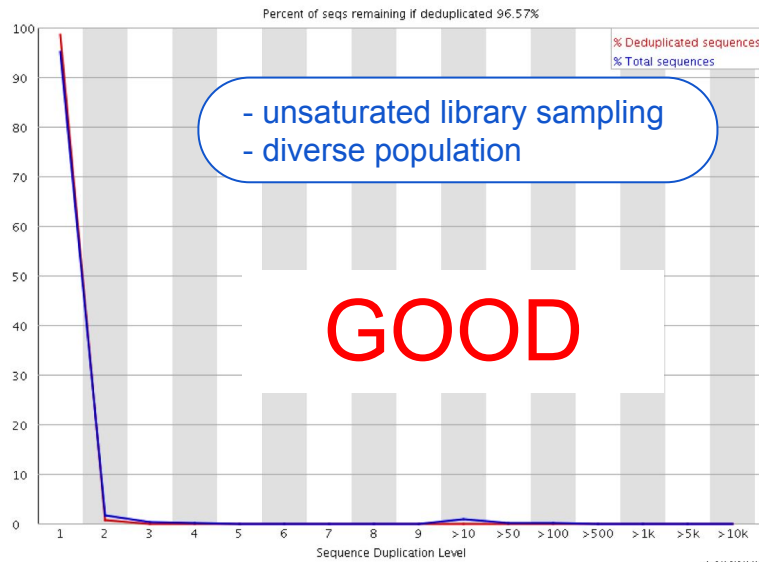
- duplication detection requires an exact sequence match over the whole length of the sequence

- biological and technical replicates can be distinguished by a system such as random barcoding

Warning: if non-unique sequences make up more than 20% of the total.

Failure: if non-unique sequences make up more than 50% of the total.

Duplication: subset enrichment



Données brutes: nettoyage

```
@SRR064166.142 HWI-EAS229_104:7:1:1:510 length=37/1
GCAAAATGGATCCGTAACCTTCGGGAAAAGGATTGGCT
+
BB@?4@B@BAB@6@?B6AB@;>)*3/:2BB4B#####
```

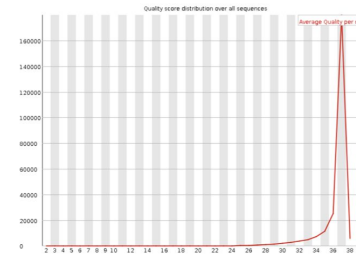
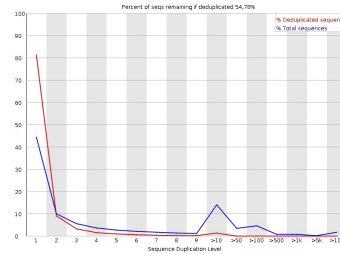
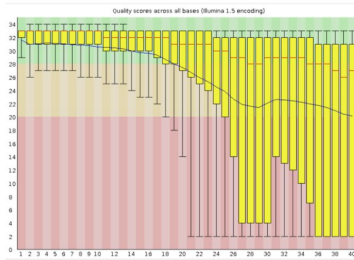
sequence identifier

sequence (IUPAC nomenclature)

sequencing quality (ASCII encoding)

Quality assessment

FastQC



Data cleaning

```
TAGCGCAATACTTTCTGTTAGCGCAAATCCTAGTAGTGCAT
CCATGTGTGGGTTGTGTTNNNNNNNNNNNNNNNNNNNNNNNN
AGTGGTATCAACGCAGAGTACGGGGGACCTTNNNNNNNNNN
```

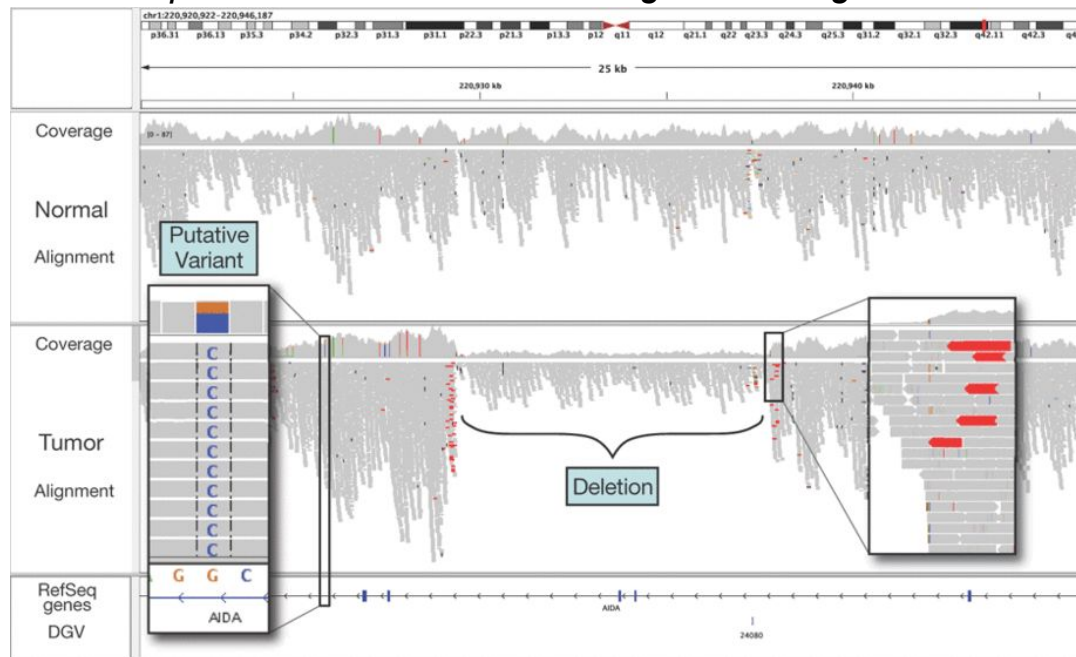
```
TAGCGCAATACTTTCTGTTAGCGCAAATCCTAGTAGTGCAT
AGTGGTATCAACGCAGAGTACGGG
```

D'où proviennent les fragments séquencés?

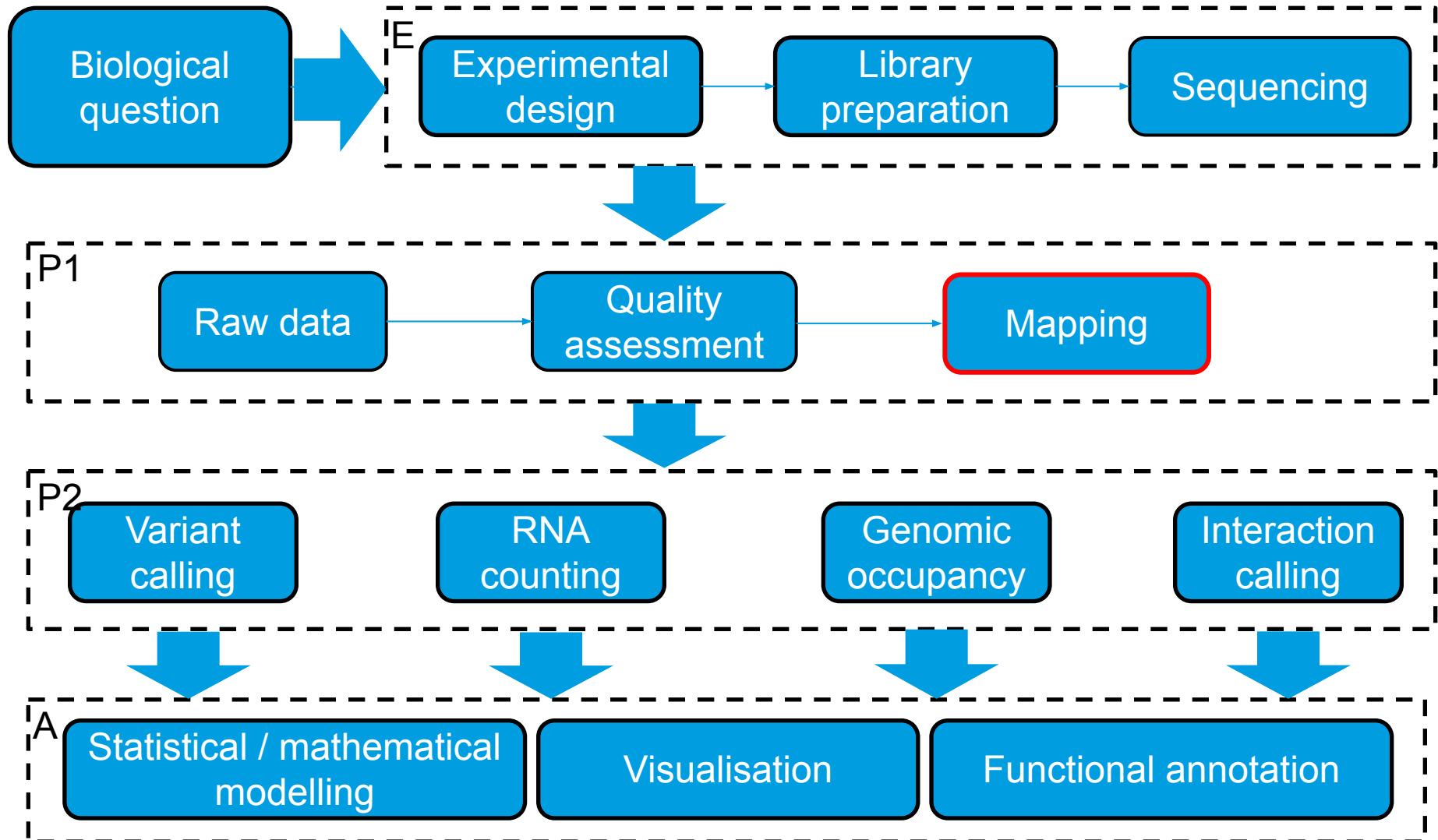
```
@SRR064166.142 HWI-EAS229_104:7:1:1:510 length=37/1
GCAAAATGGATCCGTAACCTTCGGGAAAAGGATTGGCT
+
BB@?4@B@BAB@6@?B6AB@;)>*3/:2BB4B#####
```



IGV <http://software.broadinstitute.org/software/igv/>



TD partie 1: étapes



Alignement sur référence

Alignement sur un génome de référence

Alignement

| | |
|-----------|---|
| read | CTGGTCGGATGCG |
| reference | ... GCCGGCGATGCGTCCTGGTCGGATGCGGAACGGAGCA ... |

Chercher une aiguille dans une botte de foin

Alignement

read CTGGTCGGATGCG
reference ... GCCGGCGATGCGTCCTGGTCGGATGCGGAACGGAGCA ...

Pas toujours aisé

- ❑ les reads sont courts ~100 bp
- ❑ Le génome est immense (*H. Sapien* ~3,000,000,000 bp)
- ❑ Les reads contiennent des erreurs (avec de la chance 1/1000 base est fausse)

DNA copies of
the genome



reads



assembled
genome



Quelques répétitions

Alignement

read CTGGTCGGATGCG
reference ... GCCGGCGATGCGTCCTGGTCGGATGCGGAACGGAGCA ...

Pas toujours aisé

- ❑ les reads sont courts ~100 bp
- ❑ Le génome est immense (*H. Sapien* ~3,000,000,000 bp)
- ❑ Les reads contiennent des erreurs
(avec de la chance 1/1000 base est fausse)
- ❑ Le génome contient des séquences répétées

region 1

... GCCGGCGATGCGTCCTGGTCGGATGCGGAACGGAGCAGTAAATGCCATGGAAGAGC ...

from where does it
come from?

CTGGTCGGATGCG read

region 2

... GGTTTCAGCAGGAATGCCGCTGGTCGGATGCGGAGACTCAAATGAGAACTTTGAAGGCCGAC ...

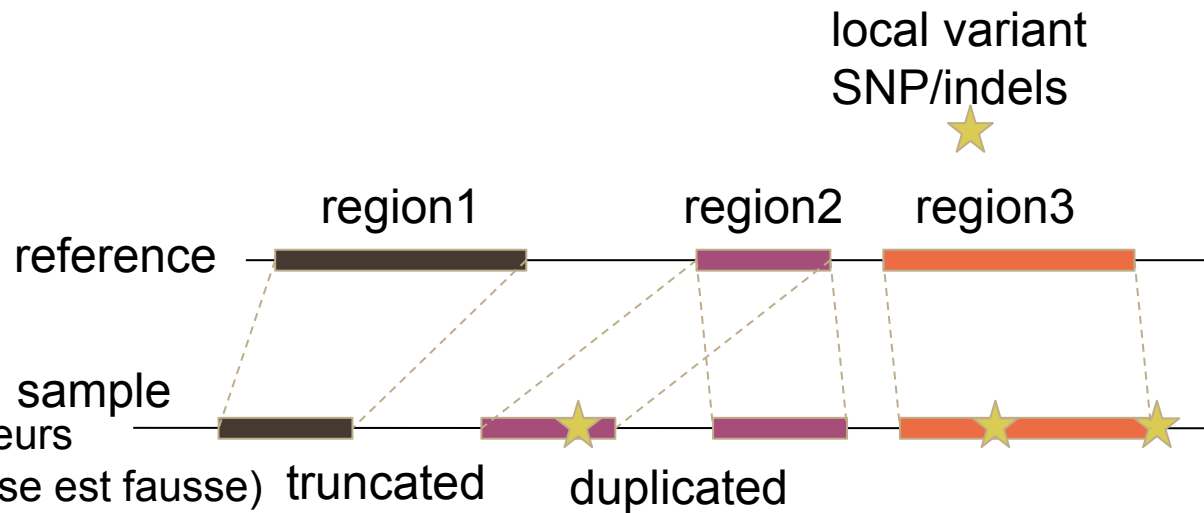
Lorsqu'on étudie des échantillons tumoraux ...

Alignement

read CTGGTCGGATGCG
reference ... GCCGGCGATGCGTCCTGGTCGGATGCGGAACGGAGCA ...

Pas toujours aisé

- ❑ les reads sont courts ~100 bp
- ❑ Le génome est immense (*H. Sapien* ~3,000,000,000 bp)
- ❑ Les reads contiennent des erreurs (avec de la chance 1/1000 base est fausse)
- ❑ Le génome contient des séquences répétées
- ❑ Variants structuraux et mutations ponctuelles



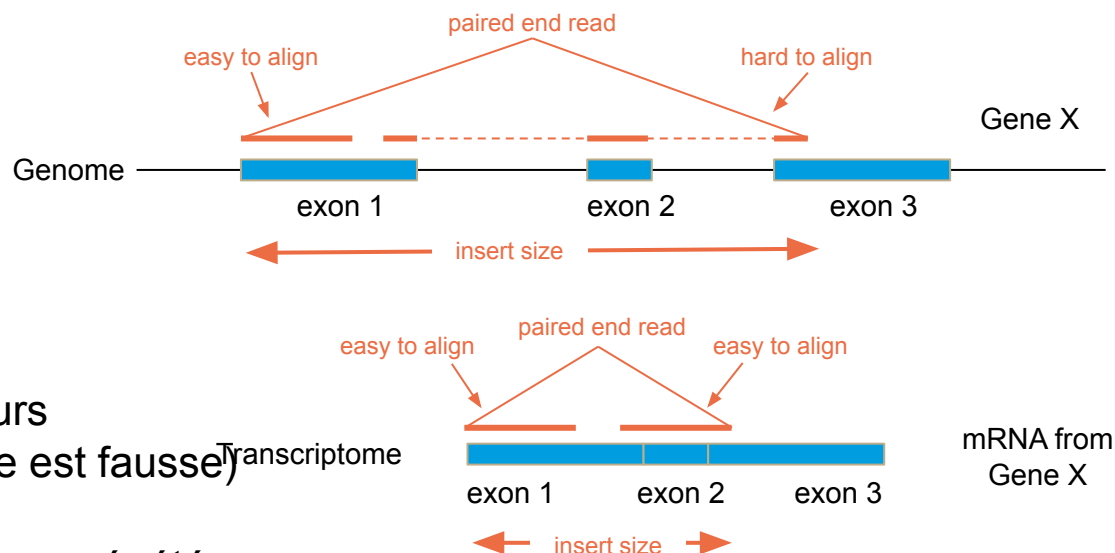
Lorsqu'on s'intéresse au transcrits ...

Alignement

read
reference ... GCCGGCGATGCGTCCTGGTCGGATGCGGAACGGAGCA ...

Pas toujours aisé

- ❑ les reads sont courts ~100 bp
- ❑ Le génome est immense (*H. Sapien* ~3,000,000,000 bp)
- ❑ Les reads contiennent des erreurs (avec de la chance 1/1000 base est fausse)
- ❑ Le génome contient des séquences répétées
- ❑ Variants structuraux et mutations ponctuelles
- ❑ Epissage



Données d'alignement: SAM / BAM format (.sam/.bam)

```
@HD VN:1.3      SO:coordinate
@SQ SN:refLN: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        *
```

| Col | Field | Type | Brief Description |
|-----|-------|--------|---------------------------------------|
| 1 | QNAME | String | Query template NAME |
| 2 | FLAG | Int | bitwise FLAG |
| 3 | RNAME | String | References sequence NAME |
| 4 | POS | Int | 1- based leftmost mapping POSition |
| 5 | MAPQ | Int | MAPping Quality |
| 6 | CIGAR | String | CIGAR String |
| 7 | RNEXT | String | Ref. name of the mate/next read |
| 8 | PNEXT | Int | Position of the mate/next read |
| 9 | TLEN | Int | observed Template LENGth |
| 10 | SEQ | String | segment SEQUENCE |
| 11 | QUAL | String | ASCII of Phred-scaled base QUALity+33 |

CIGAR: Concise Idiosyncratic Gapped Alignment Report (CIGAR) string

M: match/mismatch
 I: insertion
 D: deletion
 P: padding
 N: skip
 S: soft-clip
 H: hard-clip

Ref: GCATTCAGATGCAGTACGC
 Read: CCTCAG--GCAGTAgTg

CIGAR 2S4M2D6M3S
 POS 5

<http://samtools.github.io/hts-specs/SAMv1.pdf>

Données d'alignement: nettoyage des SAM / BAM

```
@HD VN:1.3      SO:coordinate
@SQ SN:refLN: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 *
```

| Col | Field | Type | Brief Description |
|-----|-------|--------|---------------------------------------|
| 1 | QNAME | String | Query template NAME |
| 2 | FLAG | Int | bitwise FLAG |
| 3 | RNAME | String | References sequence NAME |
| 4 | POS | Int | 1- based leftmost mapping POSition |
| 5 | MAPQ | Int | MAPping Quality |
| 6 | CIGAR | String | CIGAR String |
| 7 | RNEXT | String | Ref. name of the mate/next read |
| 8 | PNEXT | Int | Position of the mate/next read |
| 9 | TLEN | Int | observed Template LENgth |
| 10 | SEQ | String | segment SEquence |
| 11 | QUAL | String | ASCII of Phred-scaled base QUALity+33 |

Is it uniquely mapped?

Assigner les reads à des éléments génomiques

- fichiers **BAM** donnent la **localisation** des **reads alignés**
- fichiers BAM + liste d'élément génomique: compter le nombre de reads dans une région d'intérêt, les éléments sont des intervalles génomiques, coordonnées des gènes, de leurs exons, positions d'éléments régulateurs

GTF (General Transfer Format)

| | | | | | | | | |
|------|------|------|----------|----------|---|---|---|--|
| chr1 | UCSC | exon | 66999825 | 67000051 | . | + | . | gene_id "SGIP1"; transcript_id "NM_032291"; exon_number "1"; exon_id "NM_032291.1"; gene_name "SGIP1"; |
| chr1 | UCSC | 5UTR | 66999825 | 67000041 | . | + | . | gene_id "SGIP1"; transcript_id "NM_032291"; exon_number "1"; exon_id "NM_032291.1"; gene_name "SGIP1"; |
| chr1 | UCSC | CDS | 67000042 | 67000051 | . | + | 0 | gene_id "SGIP1"; transcript_id "NM_032291"; exon_number "1"; exon_id "NM_032291.1"; gene_name "SGIP1"; |
| chr1 | UCSC | exon | 67091530 | 67091593 | . | + | . | gene_id "SGIP1"; transcript_id "NM_032291"; exon_number "2"; exon_id "NM_032291.2"; gene_name "SGIP1"; |
| chr1 | UCSC | CDS | 67091530 | 67091593 | . | + | 2 | gene_id "SGIP1"; transcript_id "NM_032291"; exon_number "2"; exon_id "NM_032291.2"; gene_name "SGIP1"; |

TD partie 1: données

- Patient étudié dans le cadre de l'International Cancer Genome Consortium (ICGC) groupe Sarcome
- Données brutes (fastq, sous échantillon du chr17) provenant de:
 - échantillon tumoral (WGS et RNA-seq)
 - échantillon sain (WGS)
- Génome de référence et annotations des gènes (hg38)