Dr Issiaka SOULAMA
Pharmacien (UO)
MSc, PhD en Parasitologie (UO)
Certificat en Génétique et Génomique
(Stanford University, USA)
Mobile: 00226 70 75 71 90
00226 78 30 54 10
Centre National de Recherche et de Formation sur le Paludisme/
Institut National de Santé Publique
soulamacnrfp@gmail.com
iss.soulama@gmail.com

Plan

- Séquençage
- OMICs (Génomique-Transcriptomique-Protéomique-Métabolomique)
- Techniques de Modification du Génome
- Biotechnologie des protéines recombinantes

Le séquençage

Sommaire du Cours

- Histoire et principes de base
- Technique de Séquençage à haut débit (High-Throughput Sequencing Technology)
- Analyse de base de données de Séquençage à haut débit
- Applications du séquençage à haut débit
- ChIP-Seq
- Découvertes de l'ARN-Seq
- ENCODE : Comprendre le génome

Histoire et principes de base

Historique : Plusieurs étapes

La découverte de l'ADN

La seconde moitié du 19ème siècle: F. Miescher (biochimiste suisse) isole une substance inconnue qui contenait du phosphore, élément nouveau pour les chercheurs, et qu'elle était acide. Elle fut donc nommée acide désoxyribonucléique (ADN).

1912: fondation de la discipline de la radiocristallographie (physicien allemand, Max Von Laue)

1930: Etude de l'ADN par le britannique W. Astbury d'étudier l'ADN par radiocristallographie : découverte de la structure en long filament et de la composition par une succession de bases empilées selon un espace régulier de 0,34 nm

- **1932**: Fred Griffith un microbiologiste anglais, cherchait un vaccin contre la pneumonie, il constata que des pneumocoques virulents morts pouvaient transmettre certains de leurs caractères à des pneumocoques vivants non virulents, véhiculant ainsi la maladie.
- **1944**: Oswald Avery démontre que l'ADN était cette substance chimique qui se transmettait.

Développement de l'ingeniérie génétique

- **1965**: la découverte des enzymes dites endonucléases de restriction par W. Arber, D. Nathans et H. Smith qui rendit opératoire la biologie moléculaire. Ces enzymes avaient en effet la propriété de couper l'ADN au niveau de séquences bien spécifiques, ce qui permettait aux chercheurs d'intervenir précisément sur la molécule.
- **1971:** la première séquence comportant un gène étranger fut créée.
  - Manipulation génétique???.
  - Instauration d'un moratoire durant quelques années
  - Création d'instances de contrôle par les chercheurs puis les gouvernements optèrent

- 1983: K. Mullis invente la PCR (polymerase chain reaction), une technique permettant d'amplifier l'ADN de façon exponentielle et ainsi d'obtenir in-vitro une grande quantité d'ADN
- 1987: Introduction de la PCR en médecine légale
- A partir de 1990, les travaux de séquençage du génome de nombreuses espèces furent entrepris. On peut citer le séquençage complet du génome :
  - De virus
  - De bactéries pathogènes ou non
  - Du Nématode Coenorhabditis elegans
  - De la drosophile
  - La plante Arabidopsis thaliana

Projet du génome humain

- Programme spécifique au séquençage du génome humain lancé en 1988 sous la direction du Human Genome Organisation (HUGO) chargé de coordonner les travaux internationaux : 3 milliards d'euros furent investis à compter de 1989 pour une recherche prévue sur 15 ans
- Juin 2000: séquençage brut du génome,
- Avril 2003: Fin du projet

Rappels des principes de base

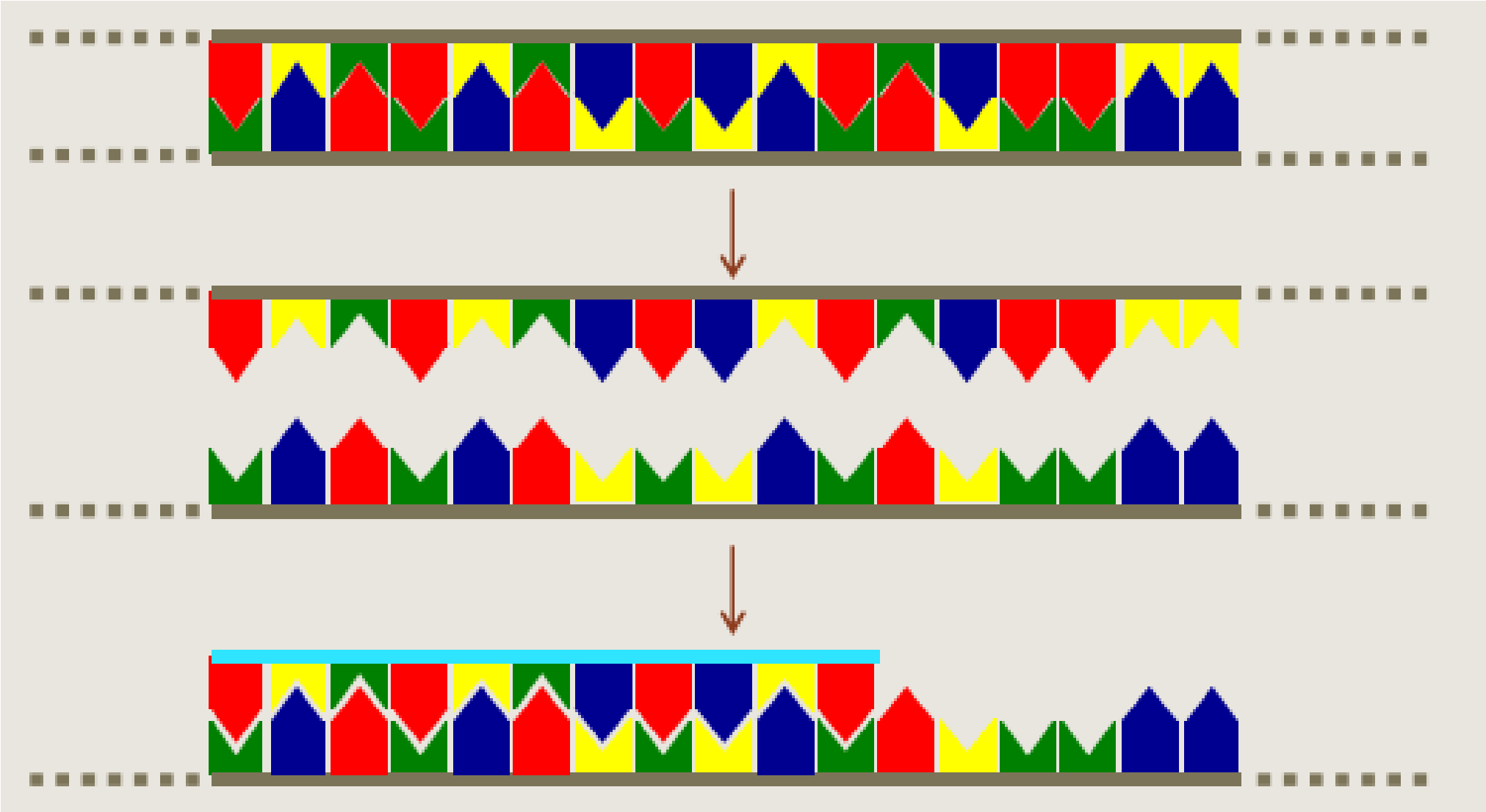
Biologie moléculaire

- Acide nucléiques: ADN, ARN
- Nucléotides
- 5' et 3
- enzyme polymérase
- PCR
- liaisons

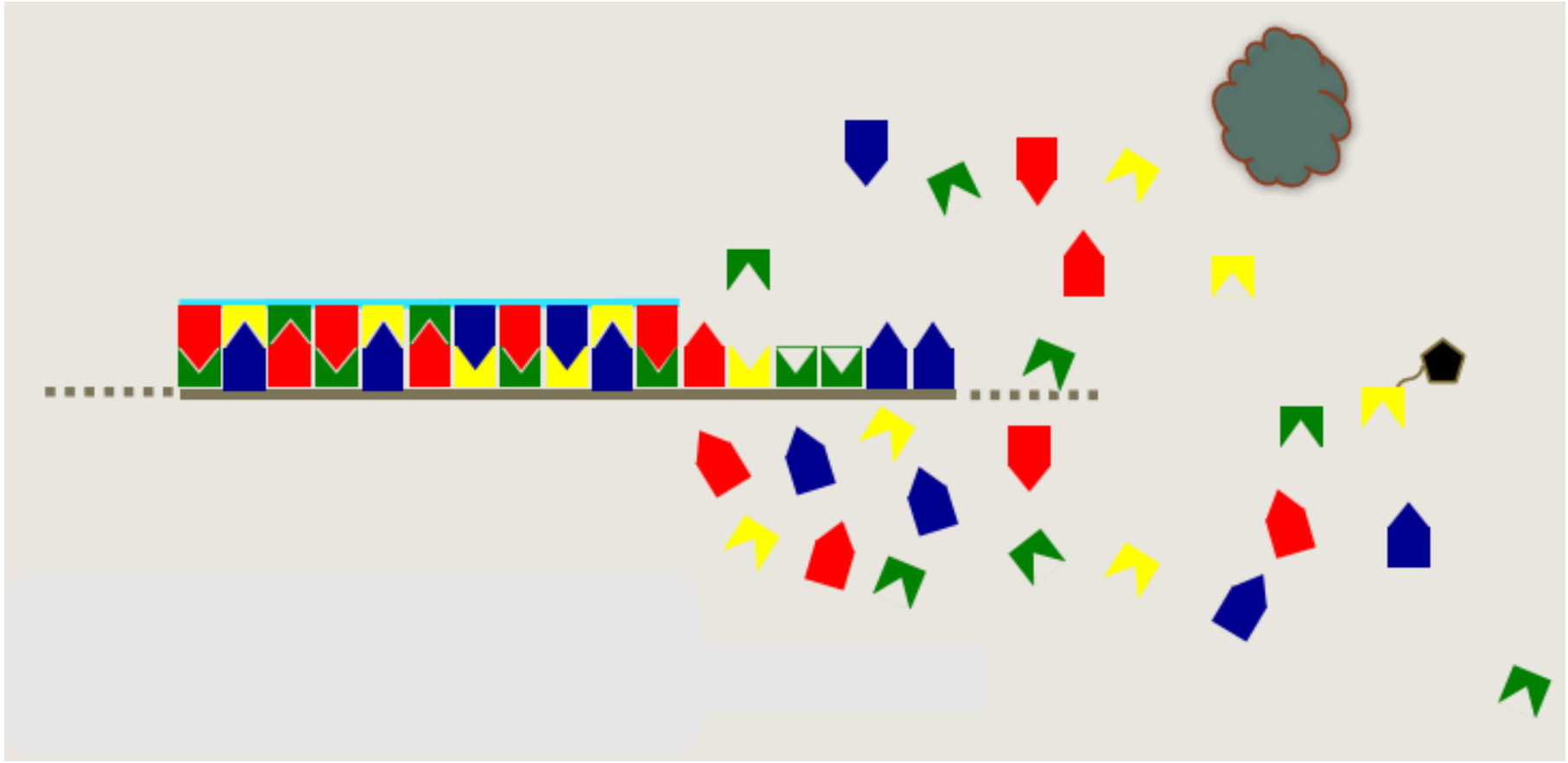
Comment séquencer l'ADN ?

- En générant des brins d'ADN et en observant comment ils sont organisés :
- Séquençage par synthèse ("SBS")
  - Utilise une ADN polymérase
  - Approche la plus courante
  - Développé par Fred Sanger ("Sanger sequencing")
  - Utilisé dans le séquençage à haut débit par : Illumina, Pacific Biosciences, Ion Torrent

Modèles de séquençage Sanger



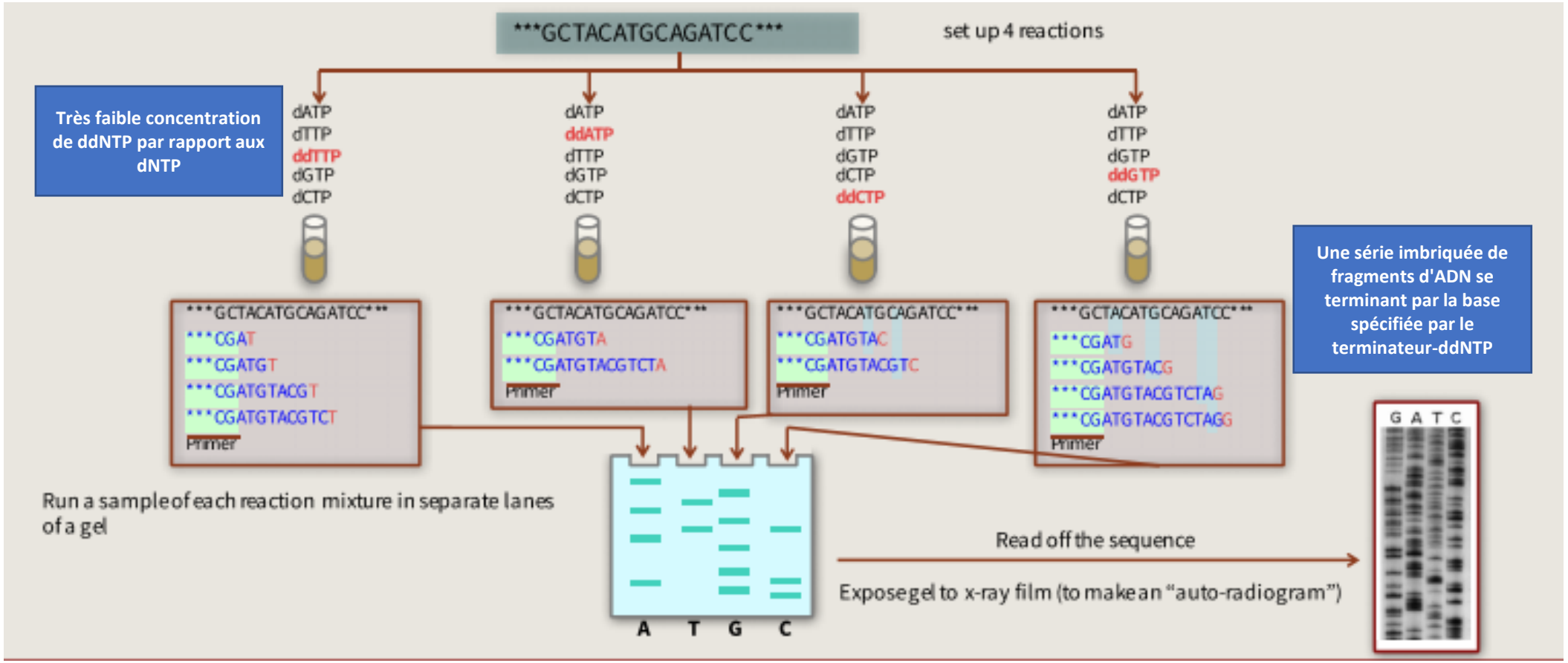
Principe du terminateur de chaîne



Le terminateur de chaîne

Les didésoxy nucléotides ne peuvent pas être rallongés davantage, et terminent donc la chaîne de séquences

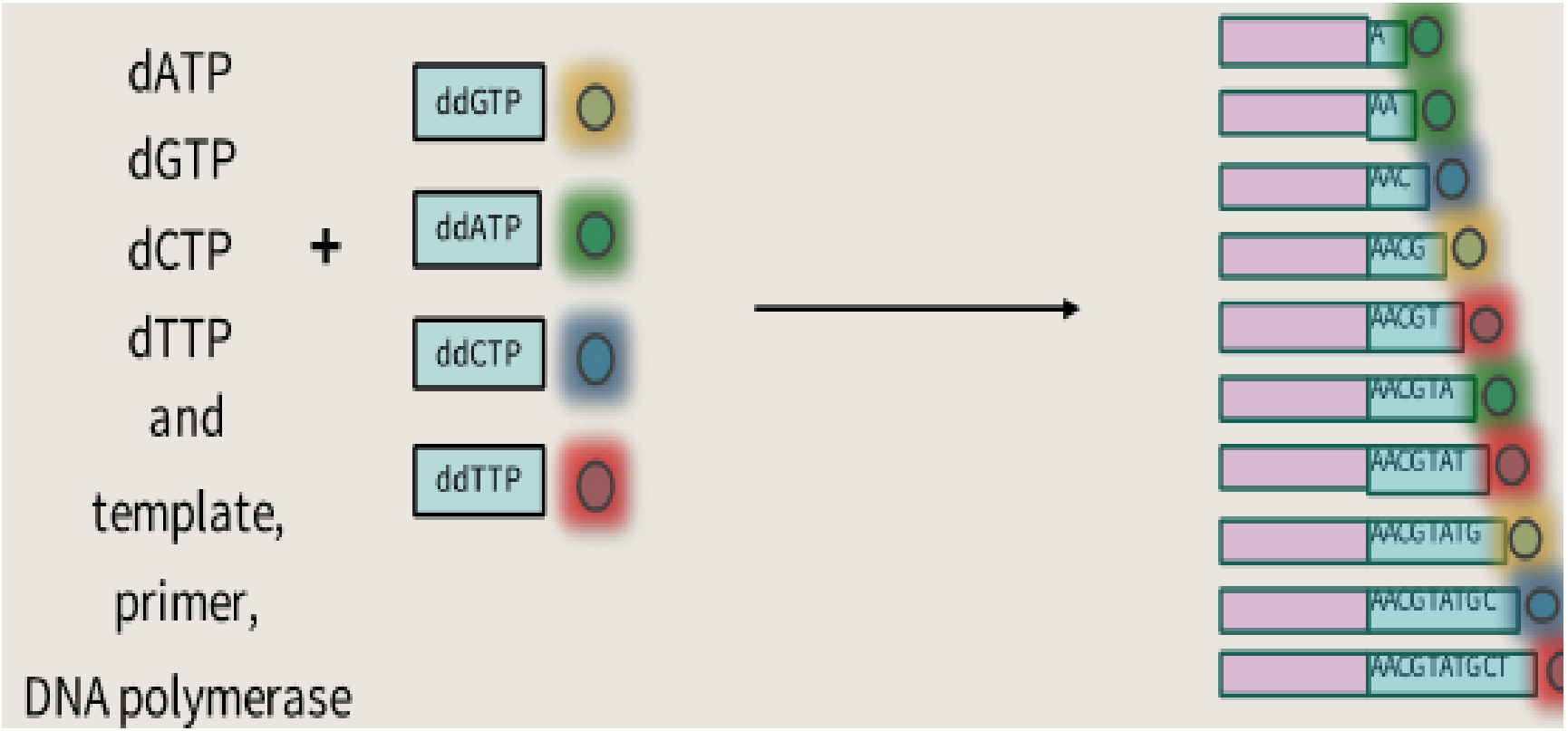
Méthode originale du Séquençage du Sanger avec un signal radioactif



- C'est super, mais...
- Ne serait-il pas formidable de tout faire fonctionner sur une seule voie ?
- Gagner de l'espace et du temps, être plus efficace
- De plus, il serait bon de tout lire au même endroit dans le gel
- Impossible de lire la séquence près du sommet, car les bandes se rapprochent de plus en plus.
- Étiquette fluorescente des ddNTP pour que chacun apparait d'une couleur différente, et peuvent être lues par un laser à un point fixe

Image: Lu Y et al. EMBO J. 1998;17:5192-5200

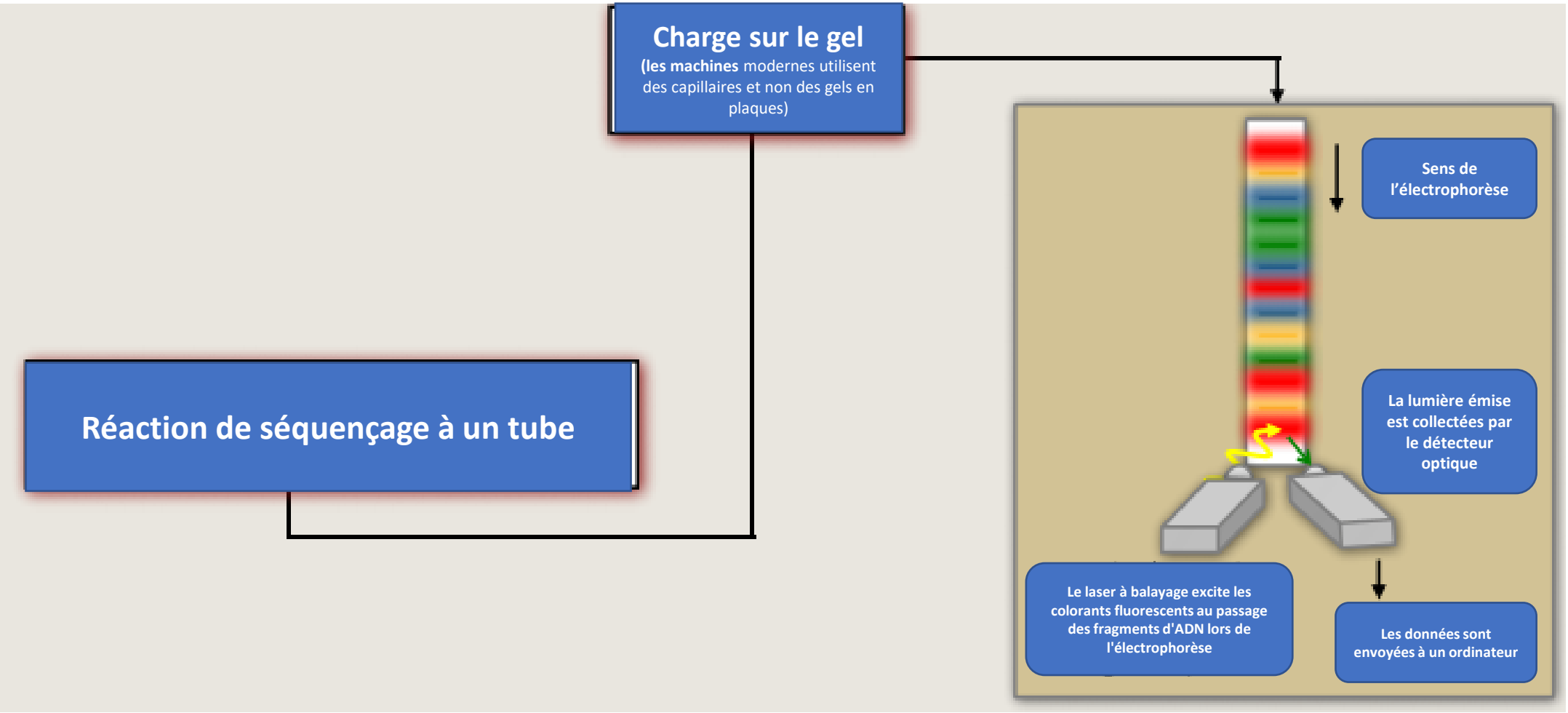
Séquençage à Fluorescence de Sanger



Echelle de séquençage dans une voie ou un capillaire sur l'appareil de séquençage de l'ADN

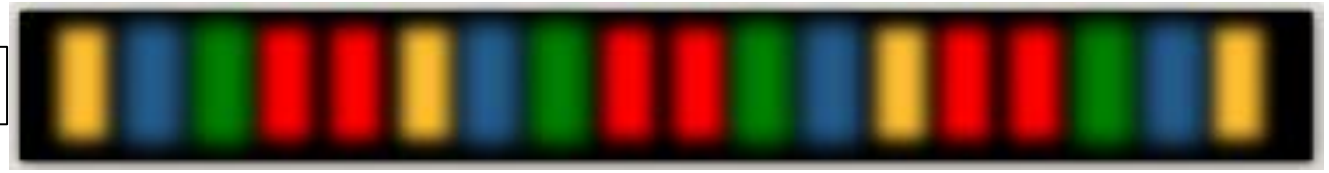
Réaction de séquençage à un tube

Séquençage à Fluorescence de Sanger



Tracé du Séquençage à Fluorescence de Sanger

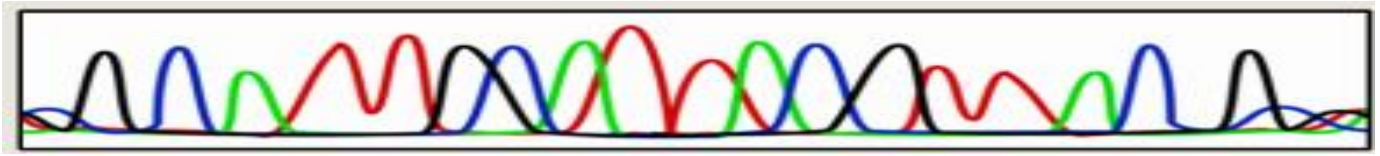
Signal fluorescent



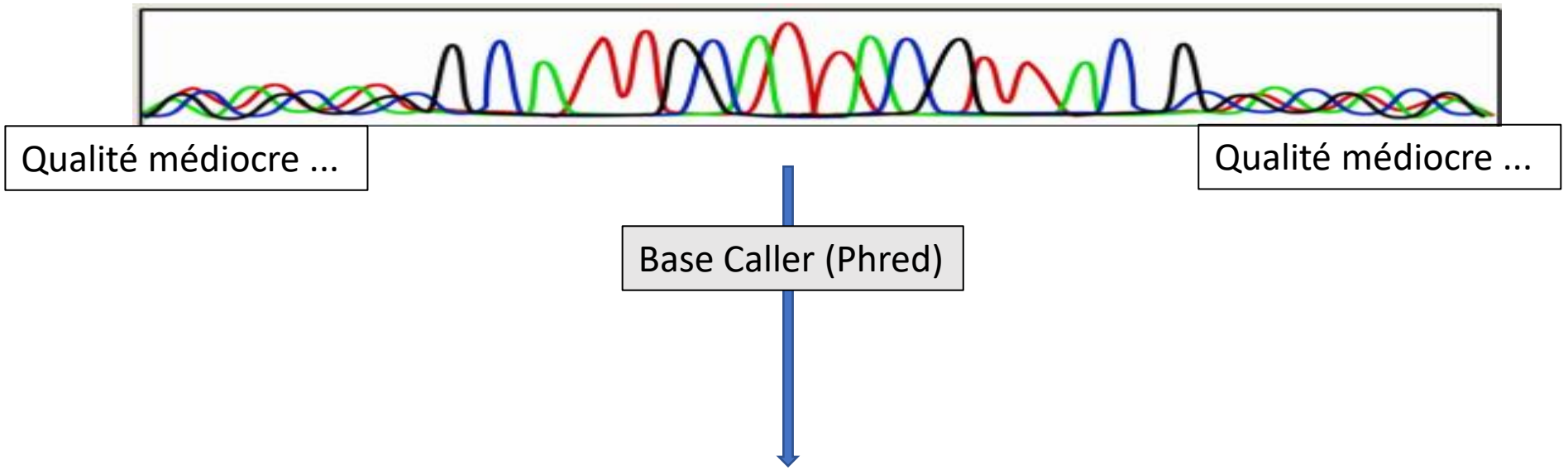
(Les signaux fluorescents réels provenant d'un gel/capillaire sont beaucoup plus laids que cela).

Divers algorithmes permettant d'augmenter le signal/bruit, de corriger les effets des colorants, les différences de mobilité, etc.

Tracé du Signal



Appel de bases




```
... 44 45 46 47 48 49 50 51 52 53 54 55 ... 718 719 720 ...
...  N  A  G  C  G  T  T  C  C  G  C  G ...   A   N   N ...
...  0  3 20 25 40 88 95 99 99 99 99 99 ...  10   0   0 ...
```

Progression de la technologie de séquençage Sanger

Gel en plaques de polyacrylamide radioactif
Faible débit, forte intensité de main-d'œuvre

Séquenceurs de gel de la gamme AB (370, 373, 377)
Séquences fluorescentes 1990-1999
6 cycles de réactions /jour
96 analyses/cycles
500 pb/fragment de lecture
288 000 bp/jour

Séquenceurs capillaires AB (3700, 3730)
1998-aujourd'hui
24 cycles de réactions/jour
96 analyses/cycles
550 - 1 000 pb/ fragment lecteur
1-2 millions pb/jour

Progression de la technologie de séquençage Sanger

Gel en plaques de polyacrylamide radioactif
Faible débit, forte intensité de main-d'œuvre

Séquenceurs de gel de la gamme AB (370, 373, 377)
Séquences fluorescentes 1990-1999
6 cycles de réactions /jour
96 analyses/cycles
500 pb/fragment de lecture
288 000 bp/jour

Séquenceurs capillaires AB (3700, 3730)
1998-aujourd'hui
24 cycles de réactions/jour
96 analyses/cycles
550 - 1 000 pb/ fragment lecteur
1-2 millions pb/jour

~Un débit multiplié par 1 000 depuis 1985 grâce à des améliorations progressives de la même technologie

Progression de la technologie de séquençage Sanger

Gel en plaques de polyacrylamide radioactif
Faible débit, forte intensité de main-d'œuvre

Séquenceurs de gel de la gamme AB (370, 373, 377)
Séquences fluorescentes 1990-1999
6 cycles de réactions /jour
96 analyses/cycles
500 pb/fragment de lecture
288 000 bp/jour

Séquenceurs capillaires AB (3700, 3730)
1998-aujourd'hui
24 cycles de réactions/jour
96 analyses/cycles
550 - 1 000 pb/ fragment lecteur
1-2 millions pb/jour

Les technologies de séquençage de deuxième génération ont un débit d'environ 500 à 30 000 fois supérieur à celui de 3730

Messages à retenir - Vocabulaire

- **Terminaison de la chaîne** : - processus chimique qui empêche les nucléotides supplémentaires de se fixer
- **Appel de base** - le processus d'attribution d'une base à un pic dans un chromatogramme
- **Phred** -Un programme informatique qui détermine la qualité d'une séquence

Technologie du Séquençage à haut débit

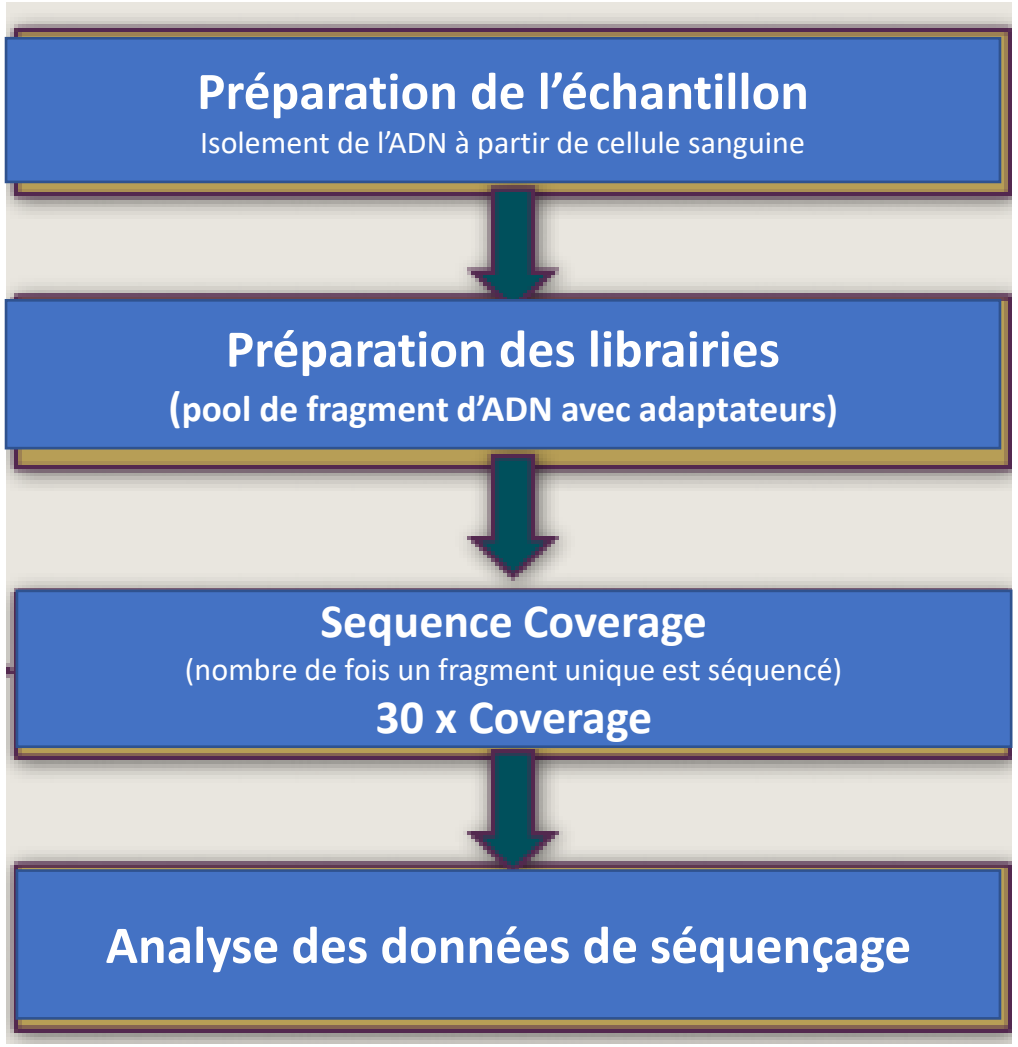
Séquençage de Sanger

- Un outil précieux pour le séquençage initial du génome
- L'achèvement du séquençage du génome humain a été le début et non la fin
- Cependant, le séquençage de millions de personnes n’est pas possible avec la méthode par Sanger - c'est trop coûteux

Différences dans le débit

Paramètres	Sanger (AB 3730)	Illumina (HiSeq 2500)
Read (fragment de lecture) Length (bp)	800	2X100
Nombre de lectures par cycle [jours]	96[<1]	6,000,000,000 [11 jours]
Débit	6Mb/jours	50 Gb/jour
Taux d'erreur SNP	Faible	élevé (~0,5%)
Taux d'erreur INDEL	Faible	Faible
Cost	**$500**	**<$0.05/Mb**

Production et analyse des données



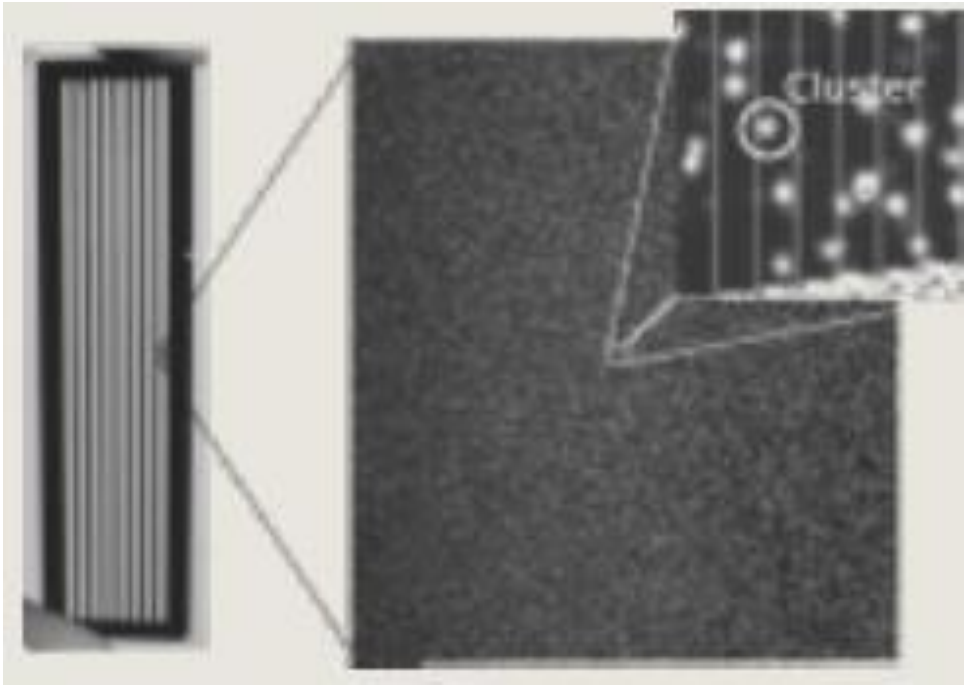
1-2 jours

10-14 jours

Jours à semaines

Le haut débit

- Séquences simultanées de millions d'ADN molécules
- « Flux cellulaire » utilisant des molécules espacées de façon aléatoire (clusters)

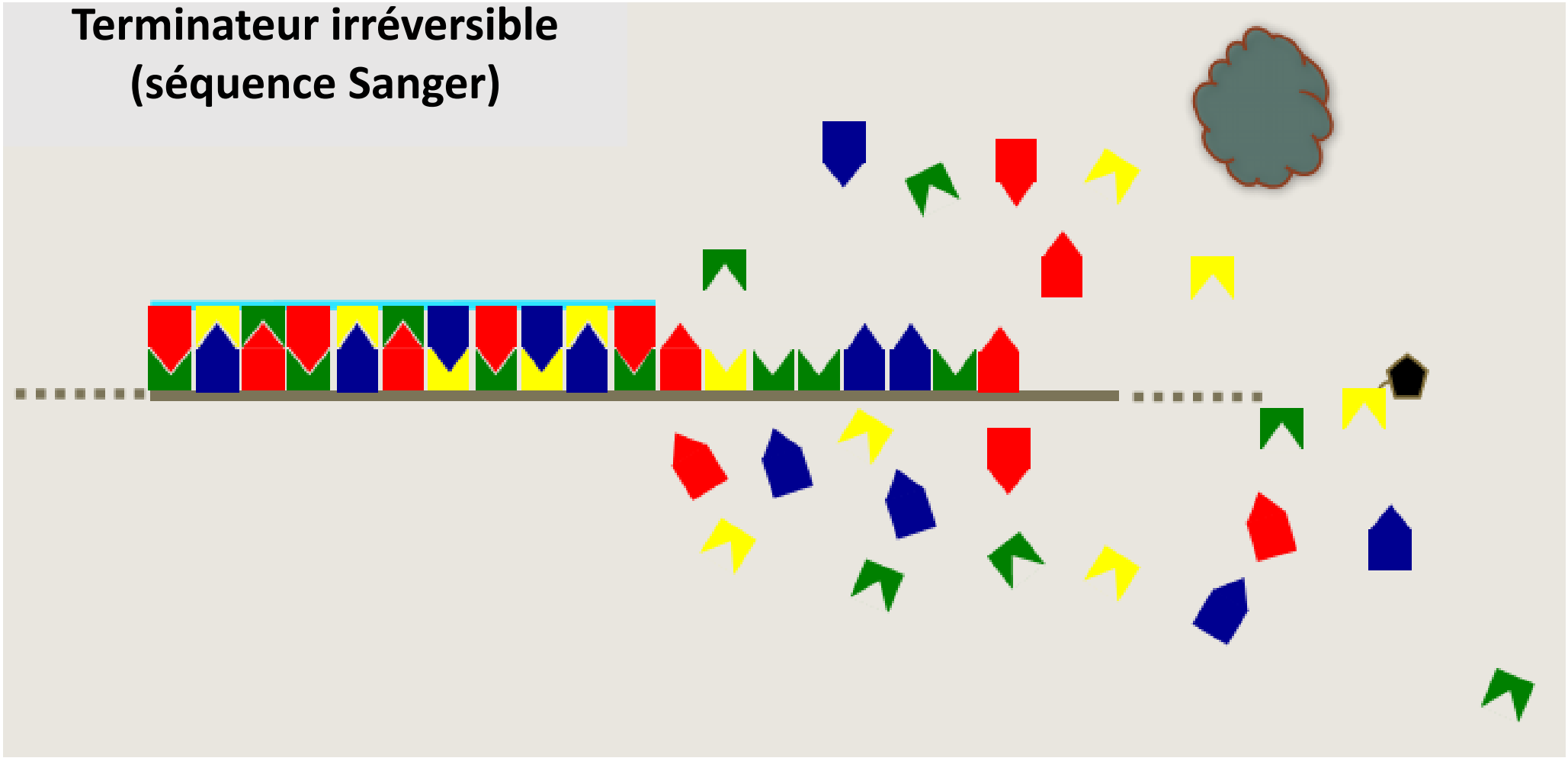


Images: Whiteford N, et al (2009). Bioinformatics25(17):2194-9

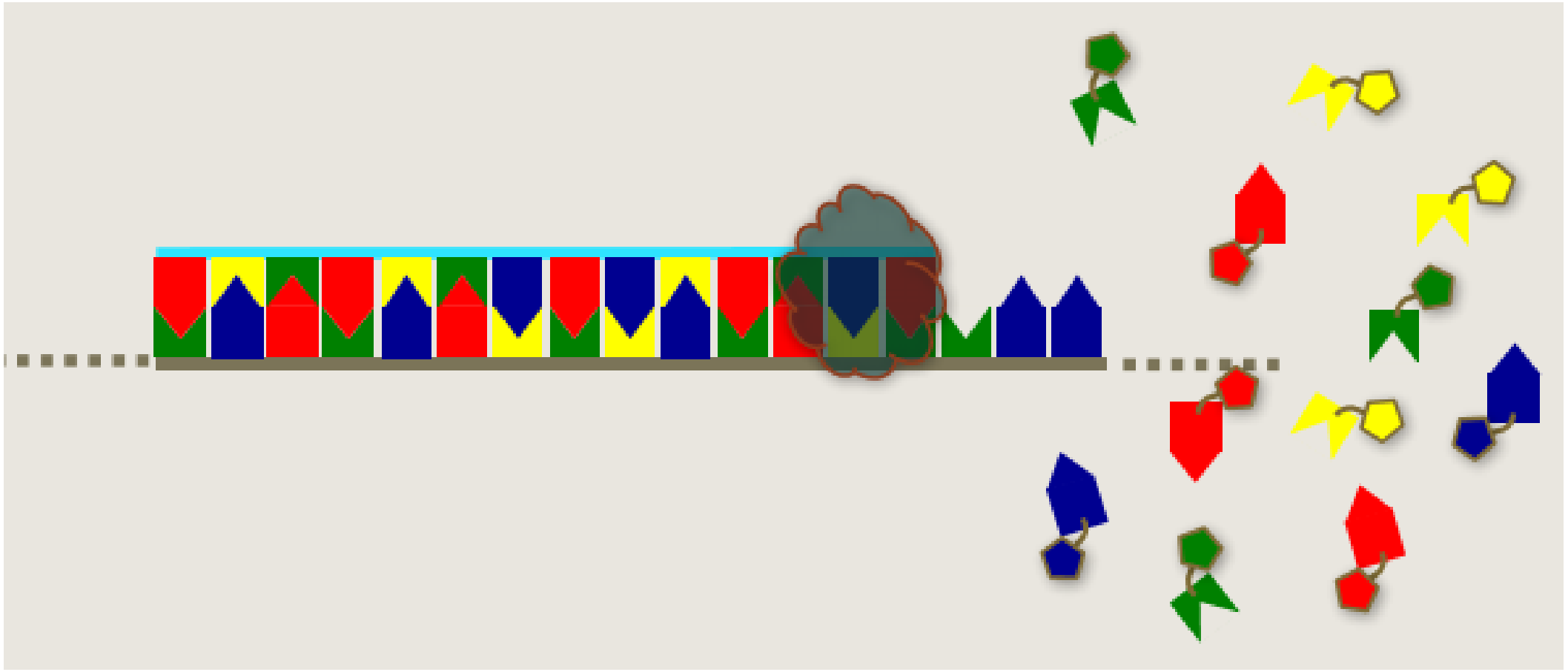
Détection

- Détection massivement parallèle sur des "colonies moléculaires" immobilisées
- Mesurer (image) chaque cycle d'addition de nucléotides, au lieu de laisser la réaction se terminer et de séparer ensuite les produits par taille (comme dans le séquençage de Sanger)
- Nécessite des nucléotides terminaux réversibles

Terminateur irréversible (séquence Sanger)

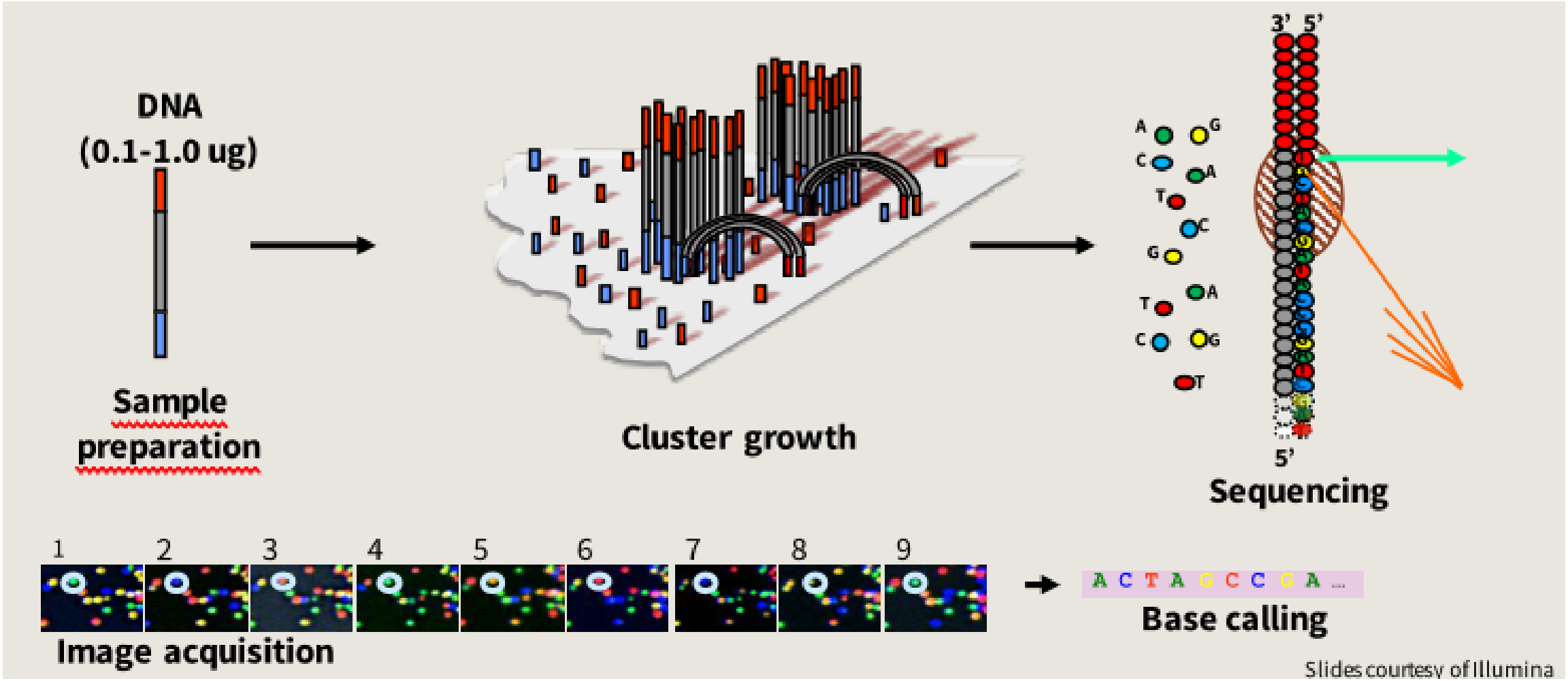


- Terminateur réversible

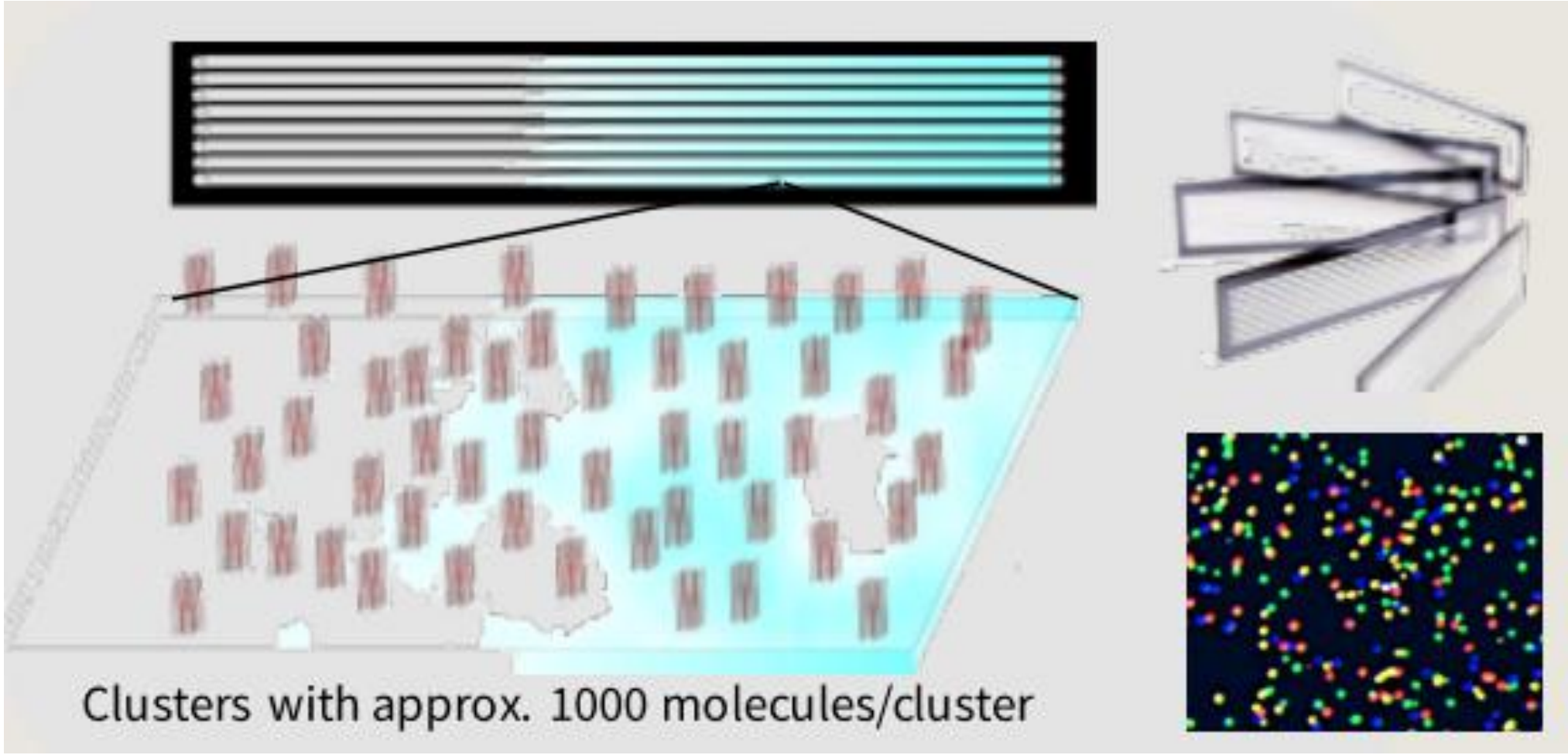


En somme

Technologie de séquençage par synthèse d'Illumina

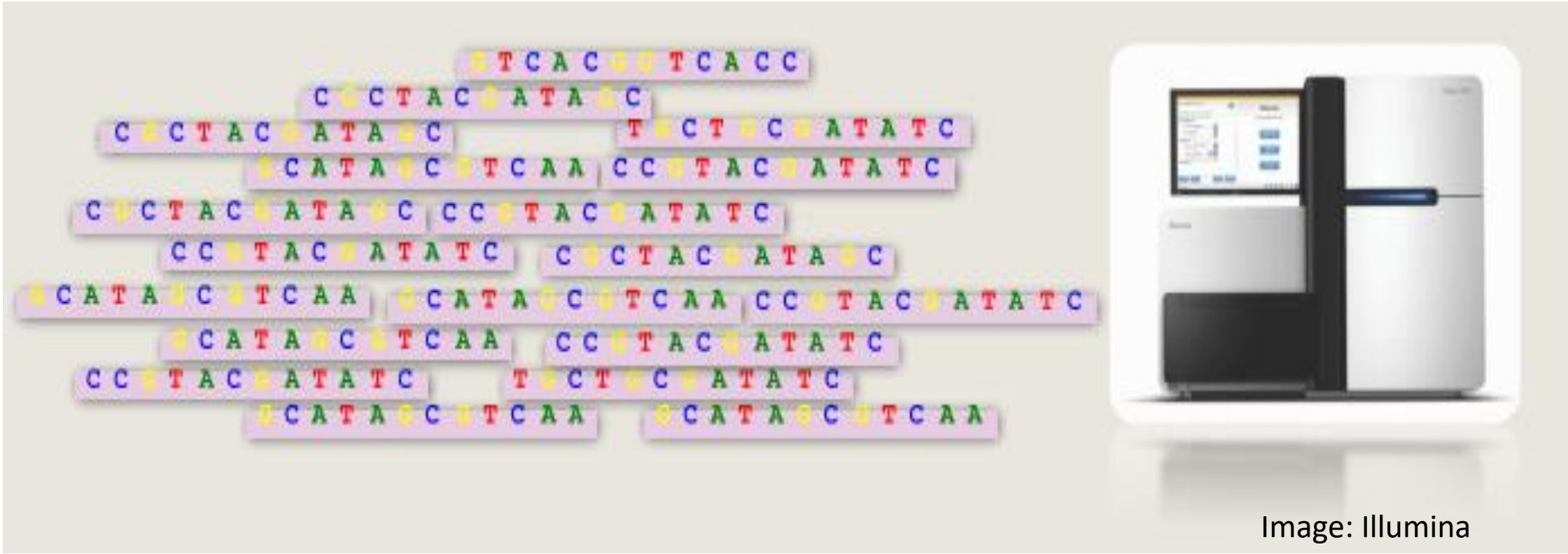


- Cellule de flux



Traitement des images, appel de bases

- Le traitement des images permet d'identifier les signaux dans les images scannées
- Le séquenceur est livré avec un ordinateur qui fait l'analyse des images pendant le séquençage



Résumé

- Read: fragment de la séquence de nucléotides
- Les améliorations de la technologie de séquençage ont conduit à des progrès majeurs dans **la collecte, la qualité et, en fin de compte, l'interprétation** des données des systèmes biologiques

Analyse des données de séquençage à haut débit

Quelles sont les données ?

- Une fois que nous avons la séquence, nous devons lui donner un sens

Comparer les segments de séquence, ou fragments de nucléotides à un génome de référence

Alignement

- AATGATACGGCGACCACCGAGATCTA

Il existe de nombreux logiciels différents qui analysent et cartographient les données de séquences

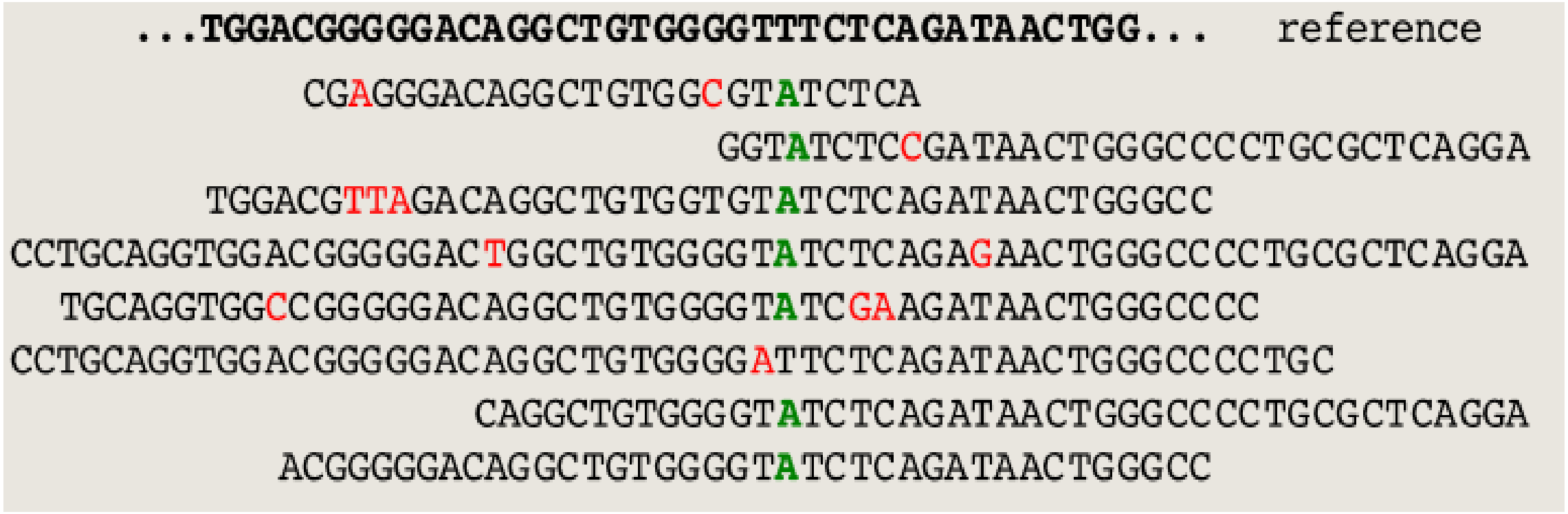
Donner un sens aux données - les lectures multiples se chevauchent

```
    ...TGGACGGGGGACAGGCTGTGGGGTTTCTCAGATAACTGG...   reference
              CGAGGGACAGGCTGTGGCGTTTCTCA
                                  GGTTTCTCCGATAACTGGGCCCCTGCGCTCAGGA
         TGGACGTTAGACAGGCTGTGGTGTTTCTCAGATAACTGGGCC
CCTGCAGGTGGACGGGGGACTGGCTGTGGGGTTTCTCAGAGAACTGGGCCCCTGCGCTCAGGA
  TGCAGGTGGCCGGGGGACAGGCTGTGGGGTTTCGAAGATAACTGGGCCCC
CCTGCAGGTGGACGGGGGACAGGCTGTGGGGATTCTCAGATAACTGGGCCCCTGC
                    CAGGCTGTGGGGTTTCTCAGATAACTGGGCCCCTGCGCTCAGGA
             ACGGGGGACAGGCTGTGGGGTTTCTCAGATAACTGGGCC
```

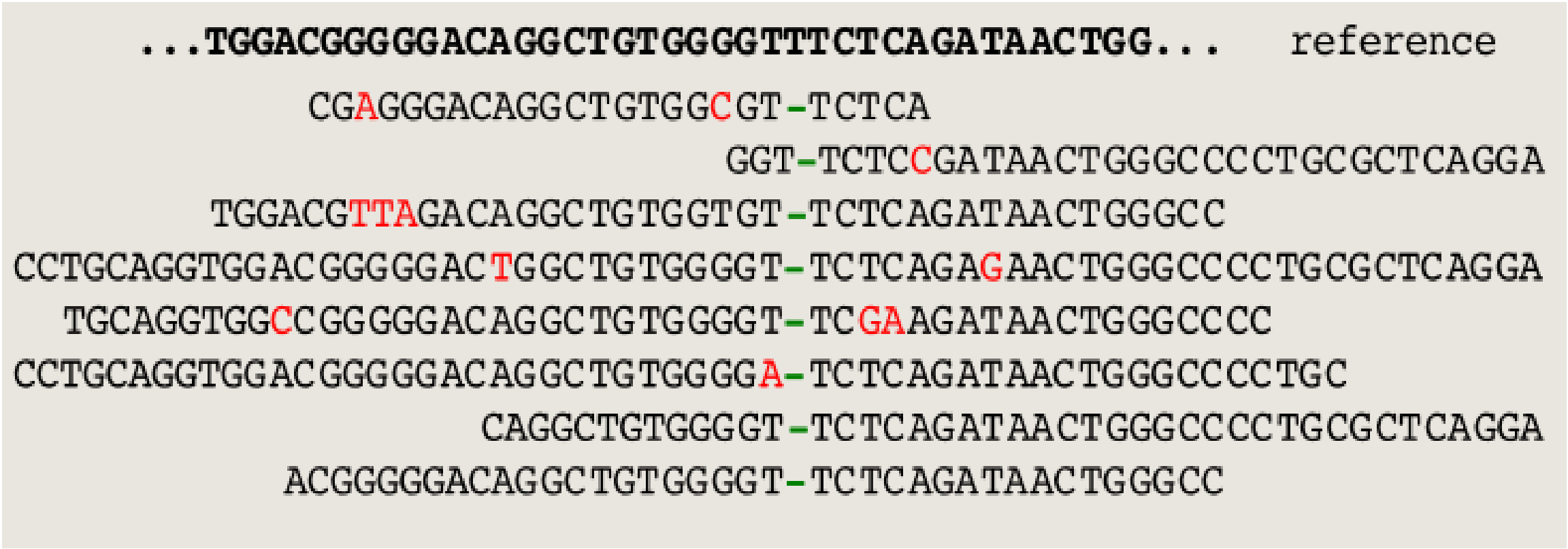
Malheureusement, les séquences de lectures ne sont jamais aussi bien assorties

Donner un sens aux données - les discordances

Identifier les SNP

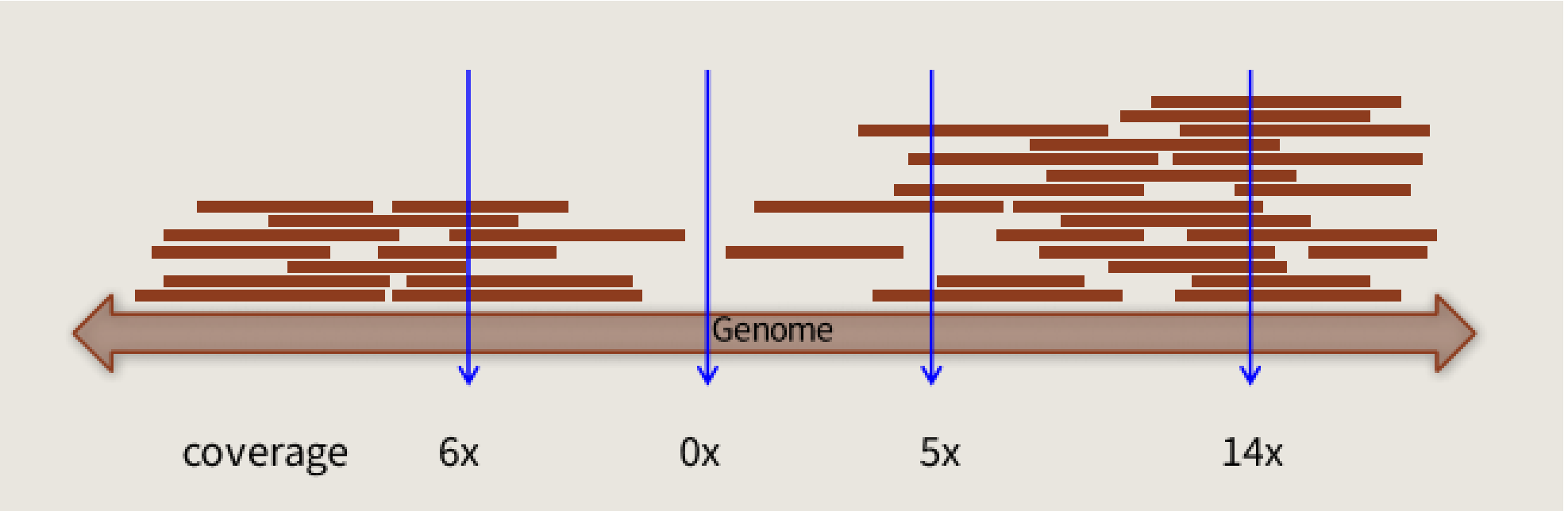


Identification des INDEL - insertions et suppressions



Couverture du séquençage

- Une couverture accrue entraîne une plus grande confiance dans l'appel de base



Une couverture accrue peut résulter de la duplication des régions du génome

Variation du nombre de copies

- La couverture de séquences d'une région peut permettre de détecter des amplifications ou des suppressions
- Le pouvoir de détection de ces régions dépend de leur taille, du nombre d'exemplaires et du nombre de lectures

Interpréter les mutations

- **Pour les SNP et les indels**
  - Est-ce qu'ils reposent dans un gène ? Si oui,
  - ont-elles un effet sur le produit protéique ?
  - Sont-elles en amont d'un gène ; affectent-elles son expression ?
- **Pour les VCN**
  - Quels sont les gènes dont le nombre de copies augmente ou diminue ?
- **Pour les variantes structurelles**
  - Quels gènes pourraient être affectés ?
  - Deux gènes sont-ils fusionnés ?

En résumé

- Séquence consensuelle - une séquence représentative basée sur le nucléotide le plus fréquent
- Contig – le consensus des fragments de séquençage qui se chevauchent
- Couverture - nombre de lectures se chevauchant sur un nucléotide particulier

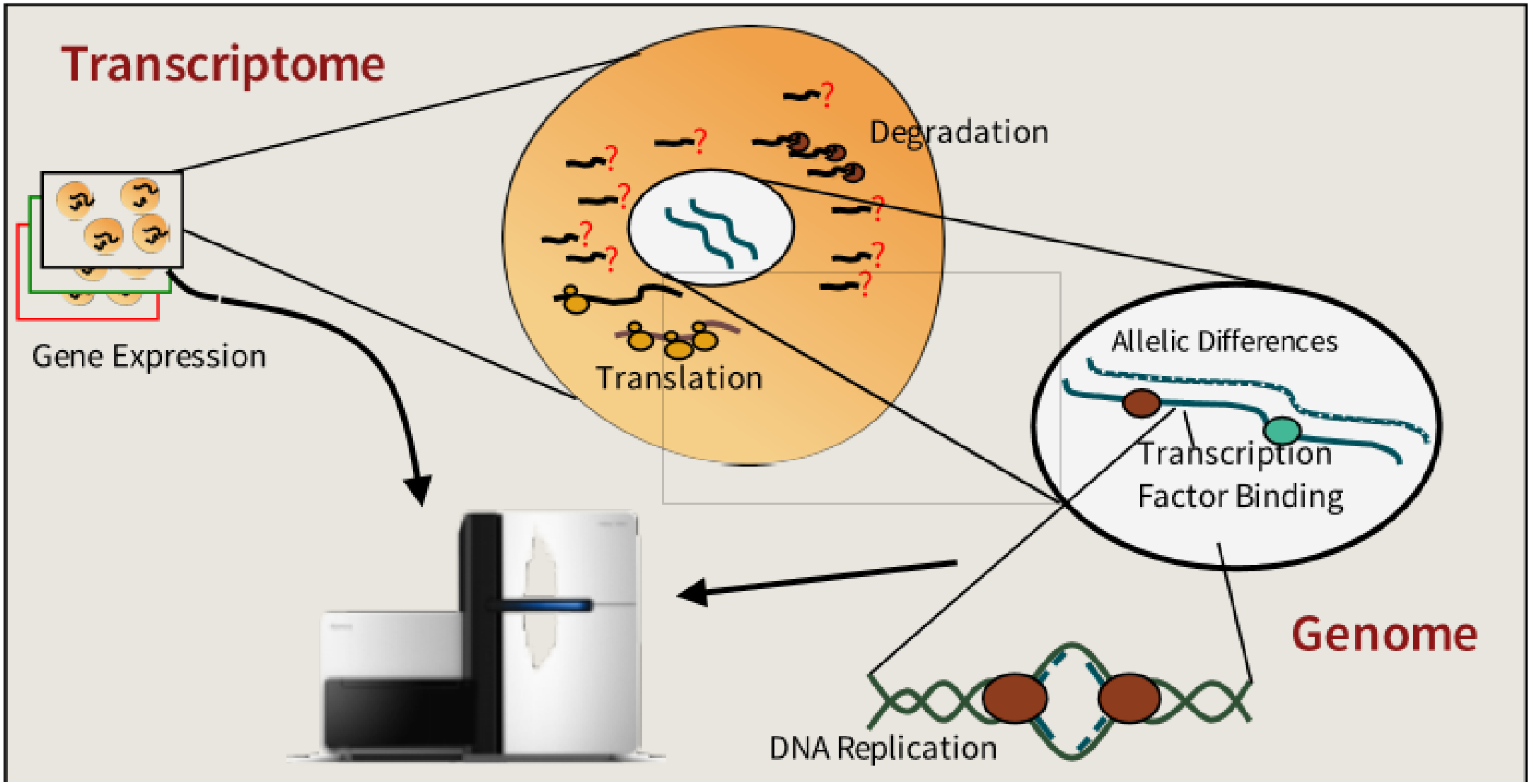
En résumé

- L'analyse des données de séquençage est l'étape qui limite le taux d'analyse du génome
- Une séquence de haute qualité est importante
- Nous pouvons identifier les mutations à partir de données de séquences à haut débit, et les distinguer des erreurs
- L'identité des mutations nous permet d'interpréter la biologie de l'échantillon que nous séquençons

Applications du séquençage à haut débit

Application du séquençage à haut débit au transcriptome

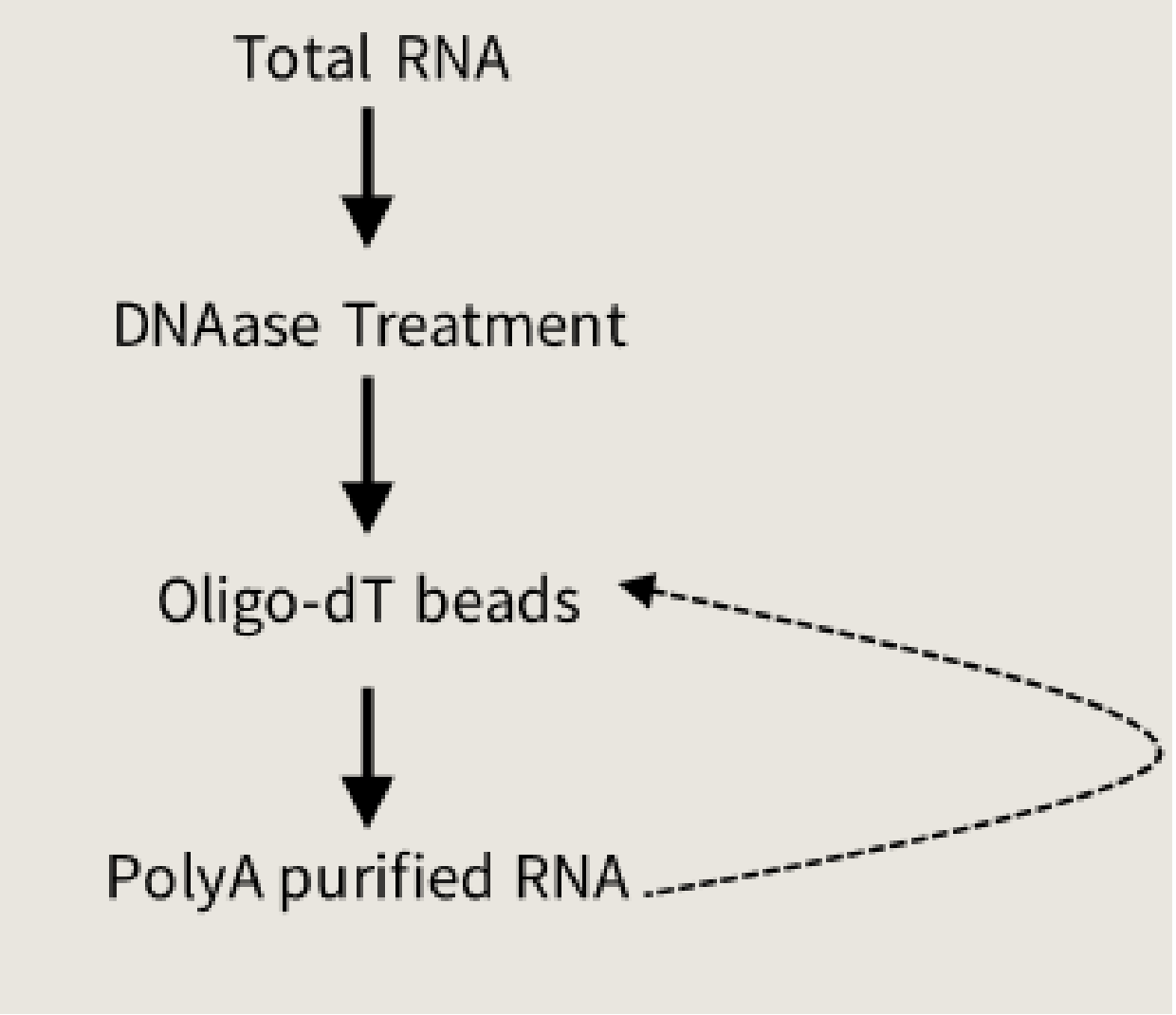
- ✓ **Qu'est-ce que l'ARN-Seq**
- ✓ **Comment générer les fragments de de séquences**
- ✓ **Comment analyser les données**
- ✓ **Comment regrouper les données d'expression**


Transcriptome
Degradation
Gene Expression
Translation
Allelic Differences
Transcription
Factor Binding
Genome
DNA Replication

ARN-seq

- Plusieurs avantages par rapport aux technologies précédentes
  - Aucune connaissance préalable n'est nécessaire pour savoir quelles parties du génome sont exprimées:
  - Permet de découvrir des sites d'épissage
  - Cartographie UTR 5' et 3
  - Différences entre les allèles
  - Facteur de transcription obligatoire
  - Découverte de nouveaux transcrits
  - Découverte de nombreuses isoformes de transcrits
  - Plus sensible

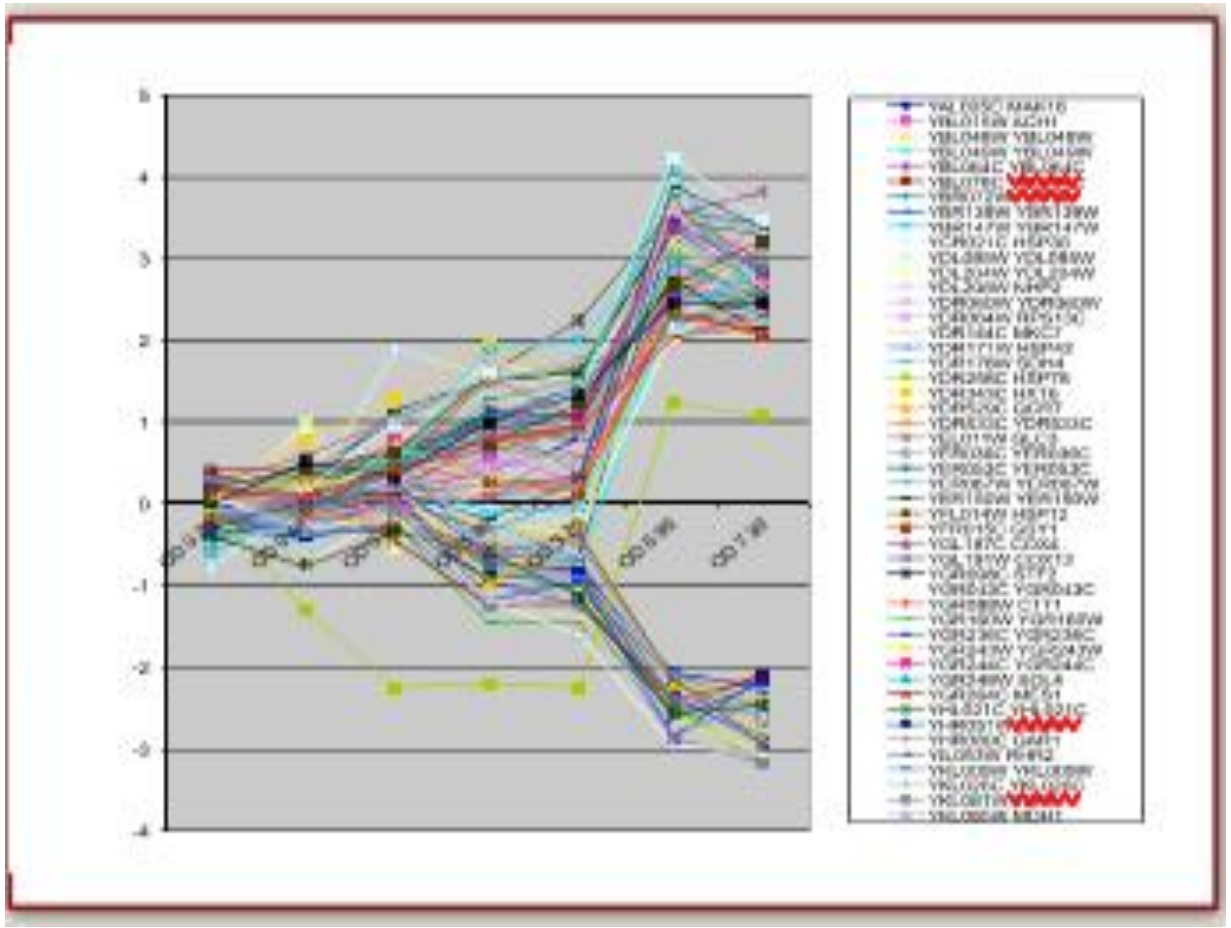
Comment séquencer l'ARNm ?

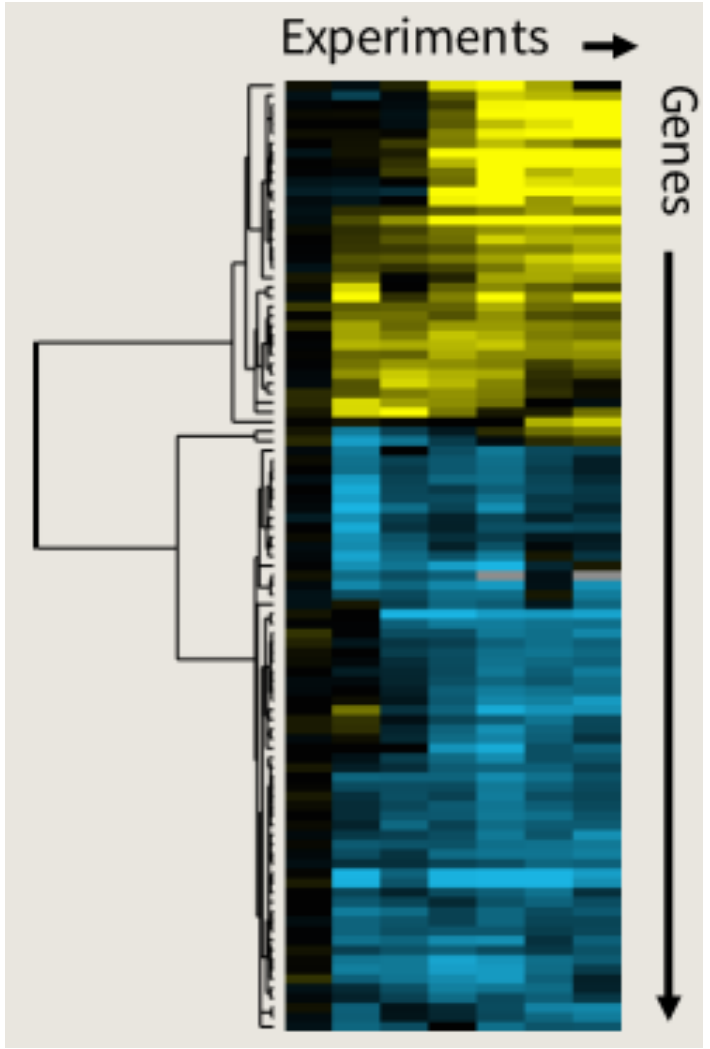


Que faire avec les données ?

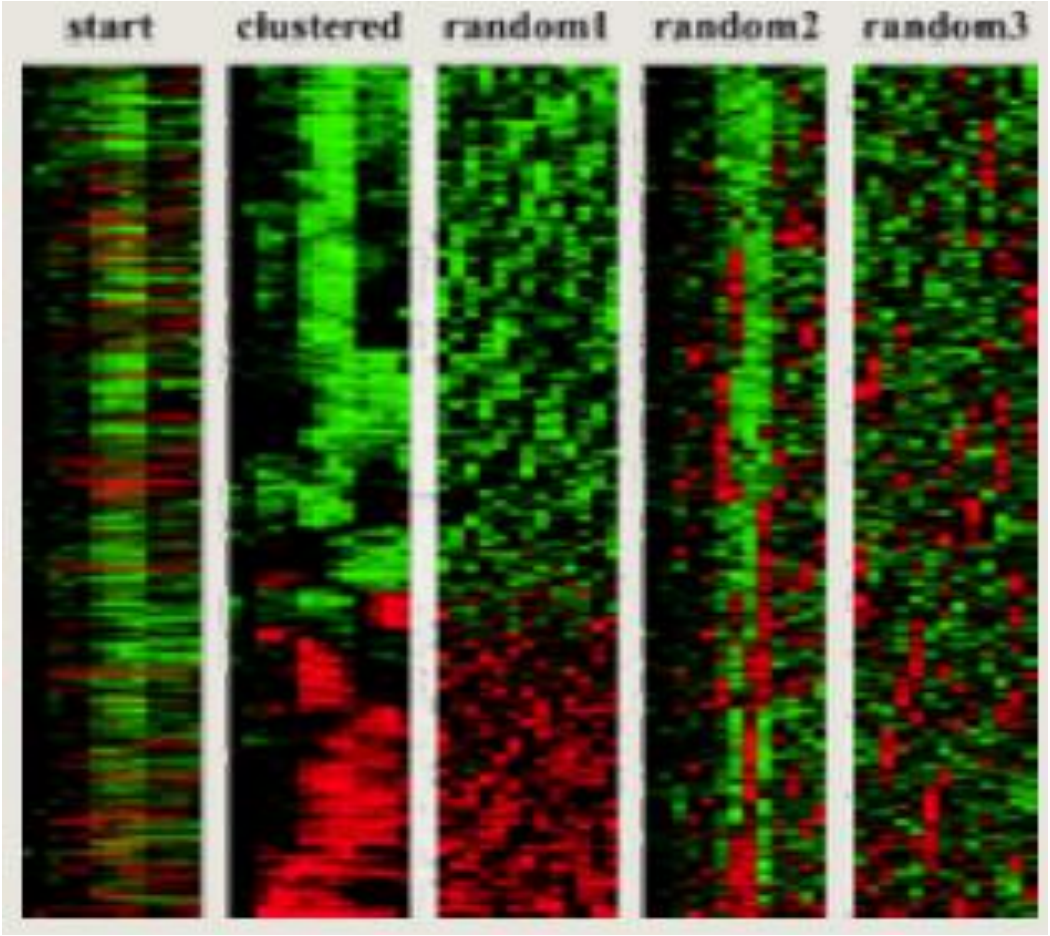
- Les données sont de courtes (50-100bp) lectures séquentielles de morceaux de transcriptions
- L'objectif est de trouver l'abondance de chaque transcription
  - Cartographier les fragments de lectures sur le génome
    - Déterminer quel est le transcrit du gène qui a généré la séquence lue
  - Pour chaque gène, compter combien de lectures correspondent à ce gène
  - Normaliser le nombre de copies en fonction de la longueur de gène
    - Parce qu'un gène plus long aura plus de fragments lus qu'un gène plus court, même s'il est exprimé exactement au même niveau
  - Normaliser le nombre de copies par le nombre total de fragments de lectures cartographiées
    - Pour que vous puissiez comparer les différentes expériences

Visualisation et Extraction des données



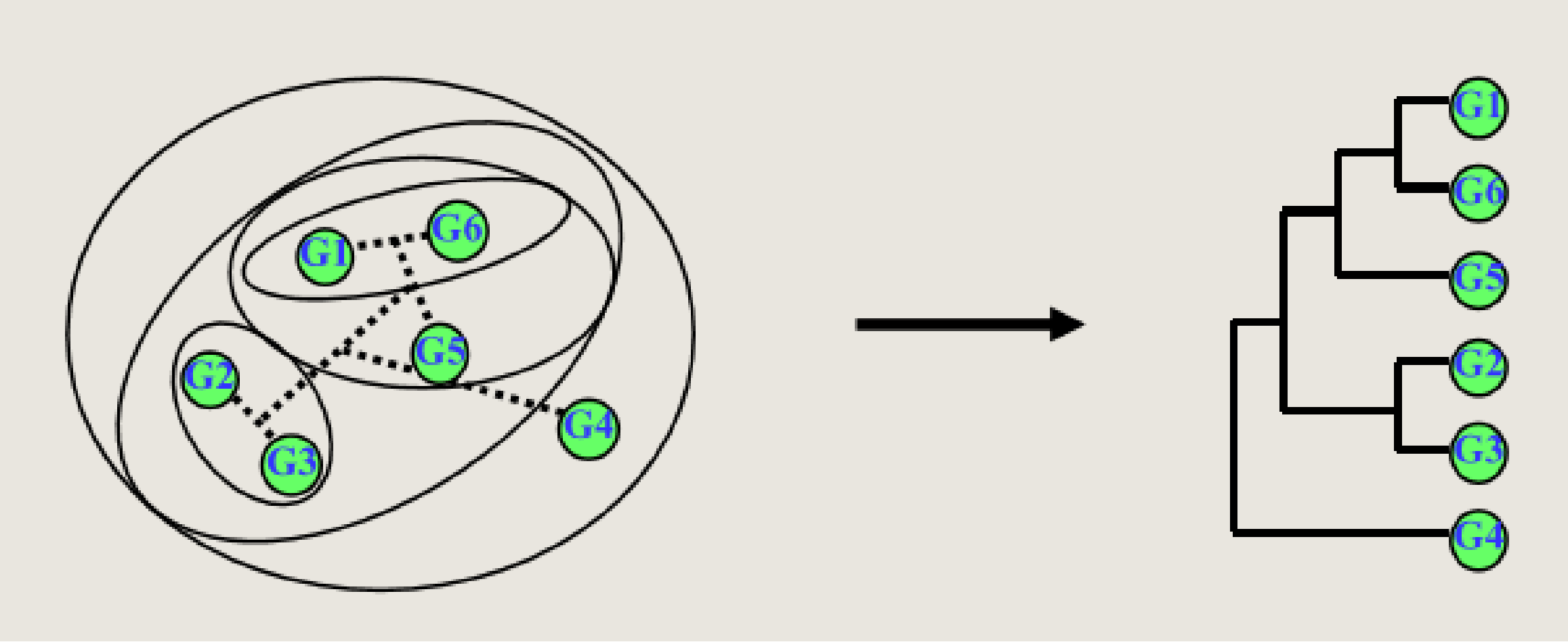


Clusters



From Eisen MB, et al, PNAS 1998 95(25):14863-8

Visualisation du regroupement hiérarchique



Résumé

- ARN-Seq - le processus par lequel nous déterminons quelles parties du génome sont transcrites, et à quels niveaux
- Le clustering - une approche qui peut être utilisée pour déterminer quels transcrits de gènes sont co-régulés