Séquençage

Bérénice Batut,
berenice.batut@udamail.fr

DUT Génie Biologique
Option Bioinformatique
Année 2014-2015

Séquençage

Séquençage ADN

- Détermination de l'ordre d'enchainement des nucléotides d'un fragment d'ADN donné

Utilité

- Comprendre et appréhender le fonctionnement des êtres vivants

Applications standards

- Assemblage de génome
- Identification de variants de structure (réarrangements) ou de séquence (SNP)
- Catalogue du transcriptome (RNA-seq)
- Mapping transcription factor binding sites (Chip-seq) or sites bound by RNA-binding proteins (CLIP-seq)
- Profilage à l'échelle du génome de marqueurs épigénétiques et structure chromatinienne (exemples, ChIP-seq, methyl-seq and Dnase-seq)

Bibliographie sur le séquençage

Hutchinson, Nucleic Acid Res, 2007

Revue sur l'histoire du début du séquençage et de la bioinformatique

Metzker, Genome Res, 2005; Metzker, Nature Rev Genet, 2010; Schadt et al, Hum Mol Genet, 2010

Revues sur les différentes méthodes de séquençage

Séquençage

Méthodes de séquençage

- 1e génération
- 2e génération
- 3e génération
- Comparaison des générations et plateformes

Traitement des données issus du séquençage

- Prétraitements des données
- Assemblage
- Analyse des séquences compilées

Séquençage

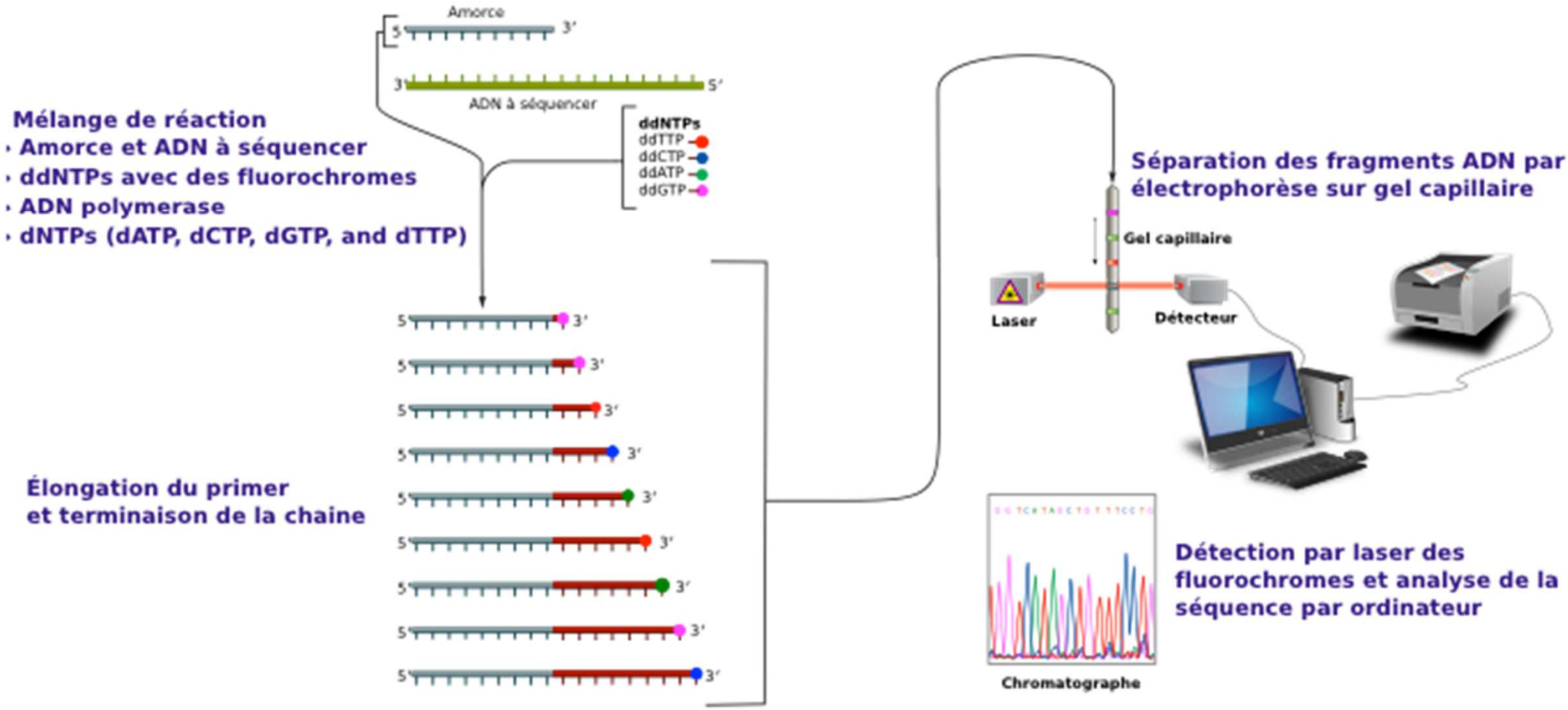
Méthodes de séquençage

- 1e génération
- 2e génération
- 3e génération
- Comparaison des générations et plateformes

Traitement des données issus du séquençage

- Prétraitements des données
- Assemblage
- Analyse des séquences compilées

1e génération des méthodes de séquençage



Méthode de séquençage de Sanger

(Sanger & Coulson, J Mol Biol, 1975; Sanger et al, Proc Natl Acad Sci USA, 1977)

Image adaptée d'une image issue de Wikipédia

1e génération des méthodes de séquençage

Avantages

- Bonne qualité de séquences jusqu'à 1kb
- 1 run = ± 2h
- Nombreux outils de traitements bioinformatique

Inconvénients

- Coût de séquençage élevé (3€ par séquence)
- Pas une méthode haut-débit (384 séquences par run)
  - Environ 10 ans et 3 milliards de dollars pour séquencer le premier génome humain

Séquençage

Méthodes de séquençage

- 1e génération
- 2e génération
- 3e génération
- Comparaison des générations et plateformes

Traitement des données issus du séquençage

- Prétraitements des données
- Assemblage
- Analyse des séquences compilées

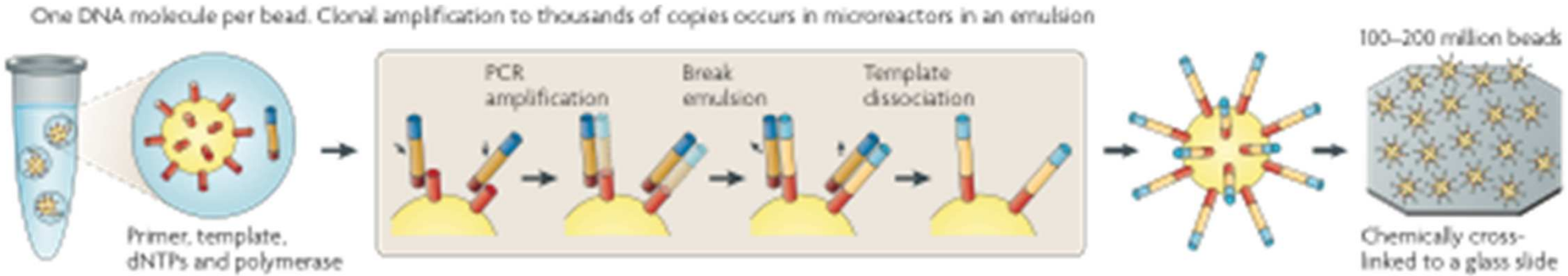
2e génération des méthodes de séquençage

- Principe
  - Préparation de l'échantillon
    - Immobilisation des templates
  - Séquençage et visualisation
    - Cyclic reversible termination (CRT)
    - Sequencing by ligation (SBL)
    - Single-nucleotide addition (SNA)
    - Real-time sequencing

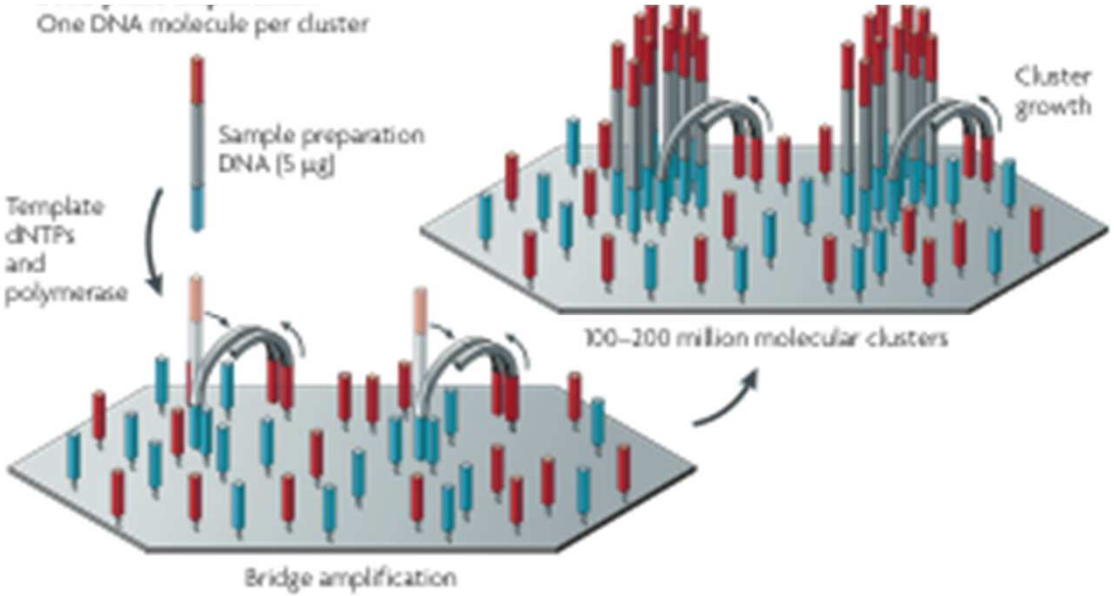
2e génération des méthodes de séquençage

- Principe
  - Préparation de l'échantillon
    - Immobilisation des templates
  - Séquençage et visualisation
    - Cyclic reversible termination (CRT)
    - Sequencing by ligation (SBL)
    - Single-nucleotide addition (SNA)
    - Real-time sequencing

Amplification clonale



Émulsion PCR

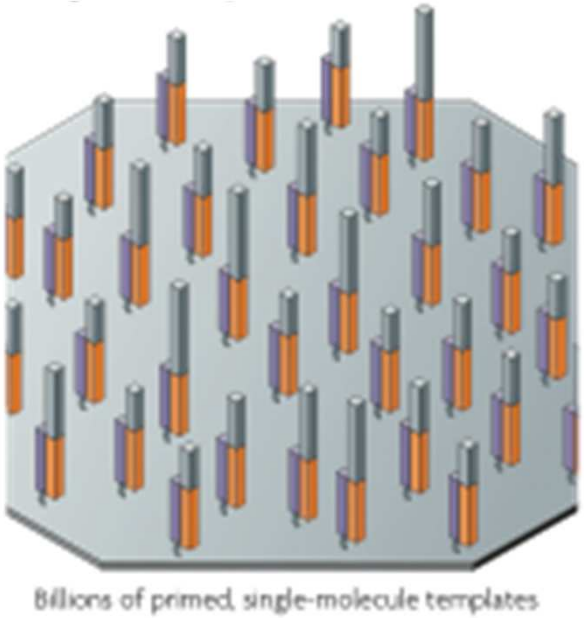


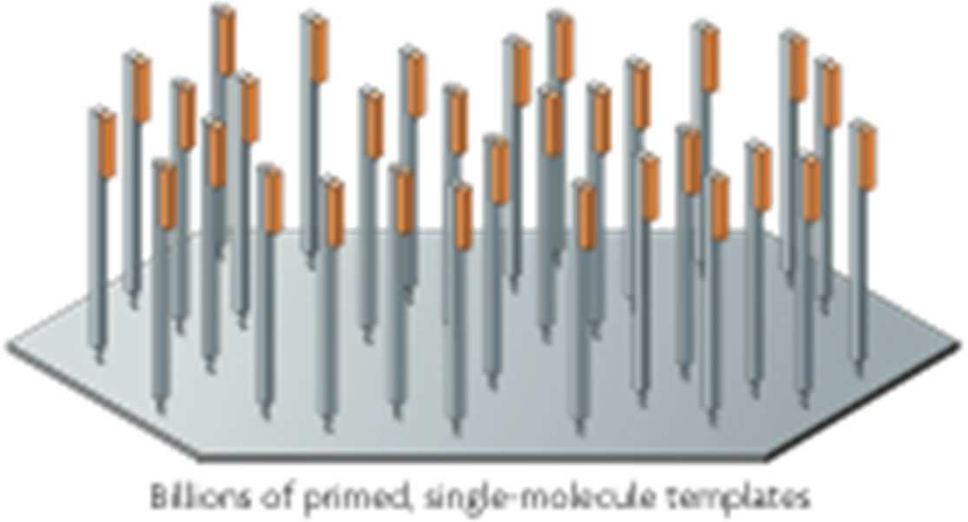
Solid-phase amplification

Images issues de Metzker, Nature Rev Genet, 2010

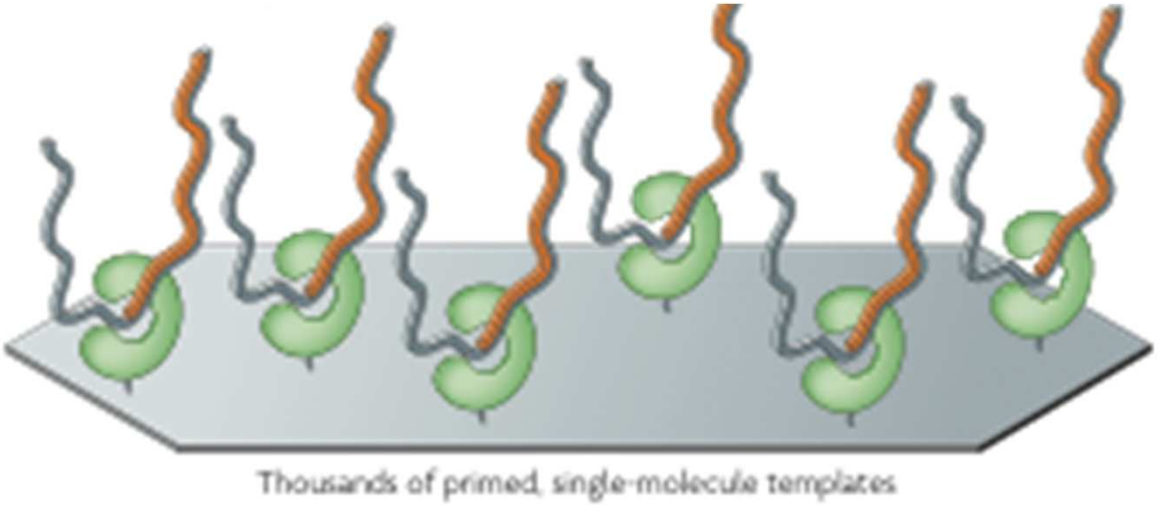
Immobilisation des templates

Immobilisation des primers





Immobilisation des polymérases



Images issues de Metzker, Nature Rev Genet, 2010

2e génération des méthodes de séquençage

- Principe
  - Préparation de l'échantillon
    - Immobilisation des templates
  - Séquençage et visualisation
    - Cyclic reversible termination (CRT)
    - Sequencing by ligation (SBL)
    - Single-nucleotide addition (SNA)
    - Real-time sequencing

Cyclic reversible termination (CRT)

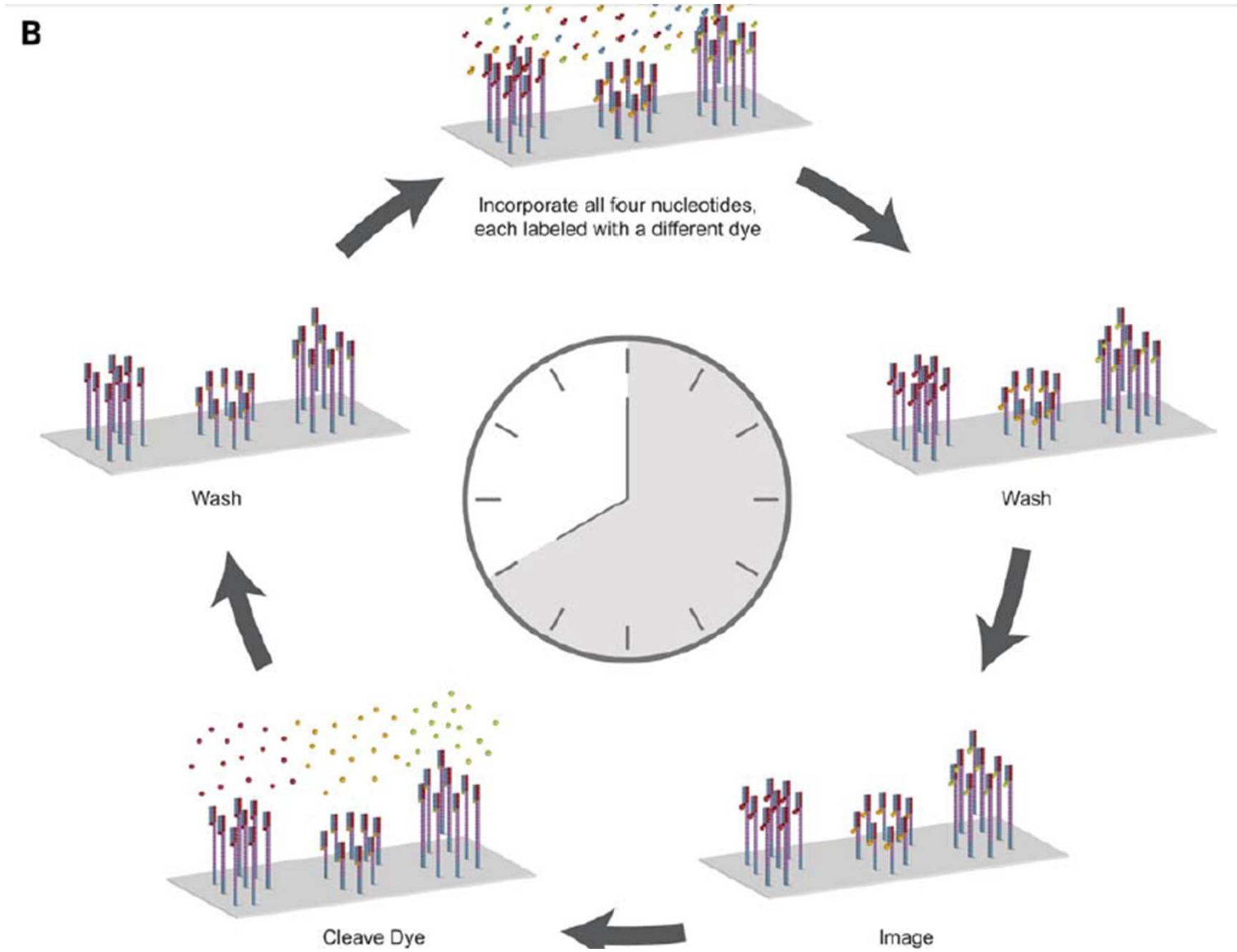
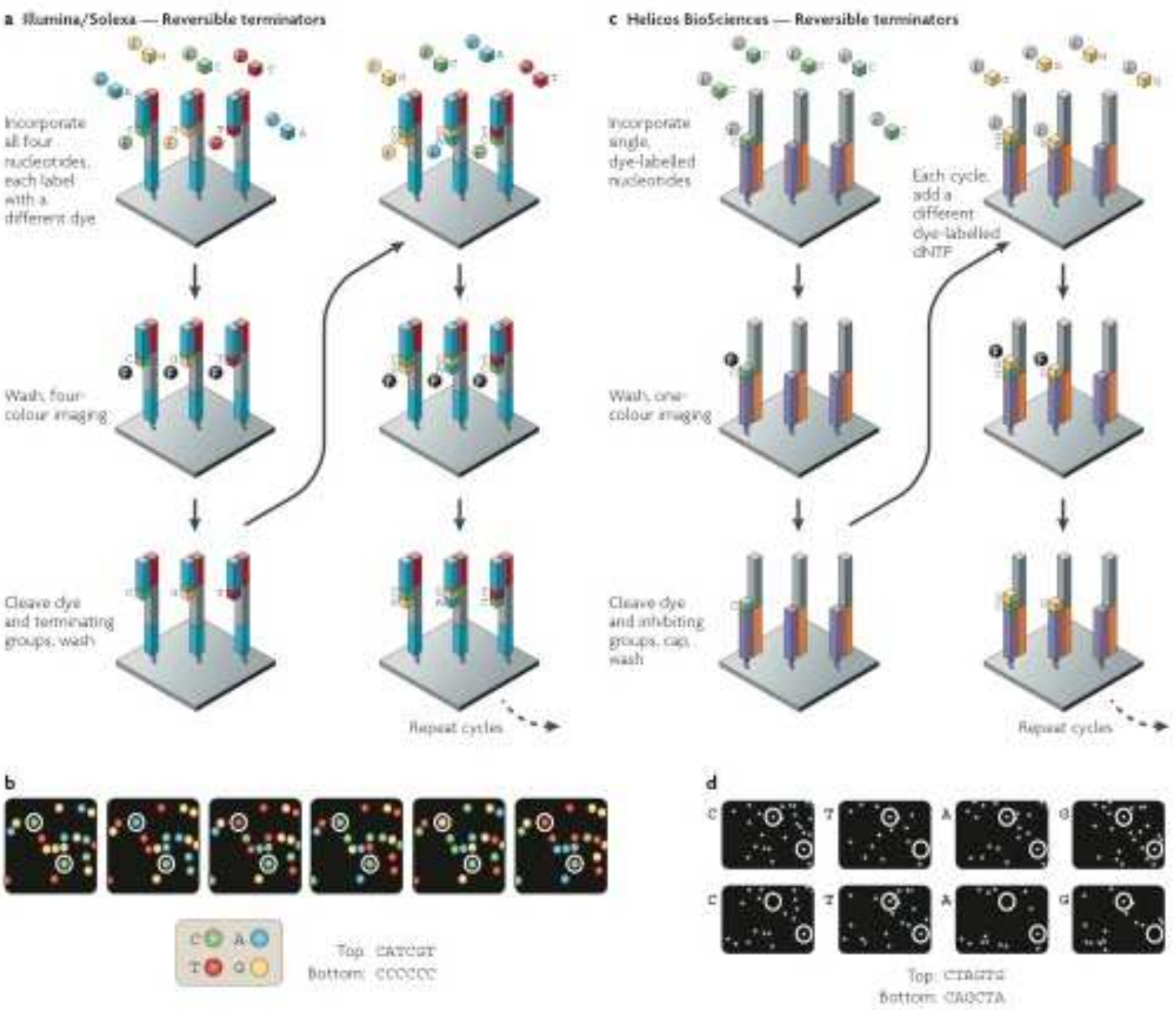


Image issue de Schadt et al, Hum Mol Genet, 2010

Cyclic reversible termination (CRT)



Images issues de Metzker, Nature Rev Genet, 2010

2e génération des méthodes de séquençage

- Principe
  - Préparation de l'échantillon
    - Immobilisation des templates
  - Séquençage et visualisation
    - Cyclic reversible termination (CRT)
    - Sequencing by ligation (SBL)
    - Single-nucleotide addition (SNA)
    - Real-time sequencing

Sequencing by ligation (SBL)

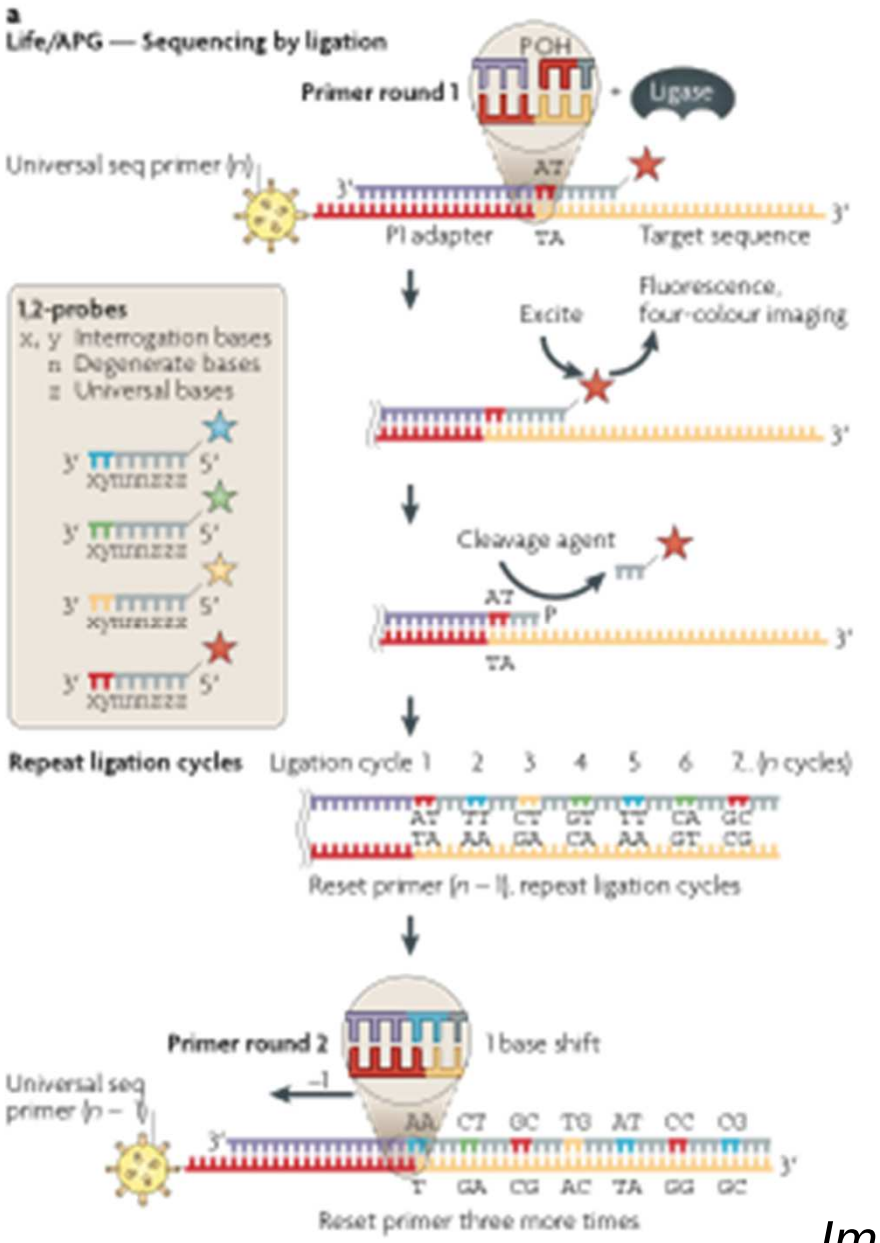
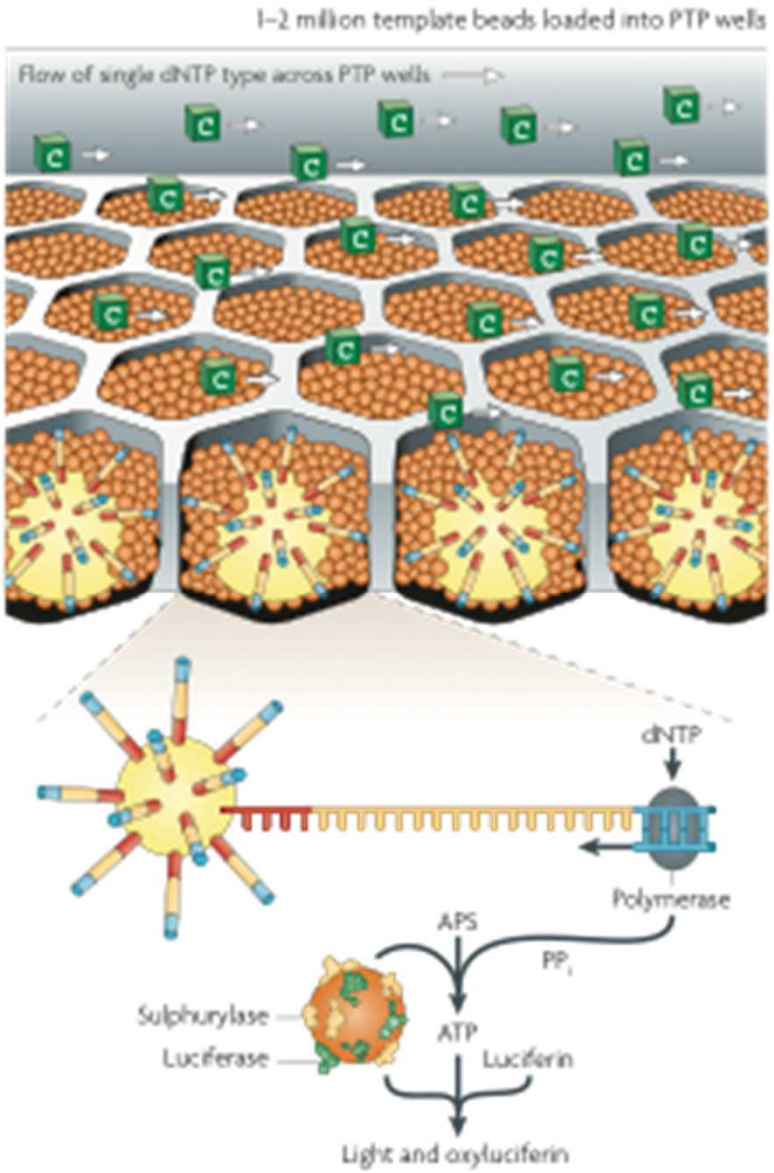


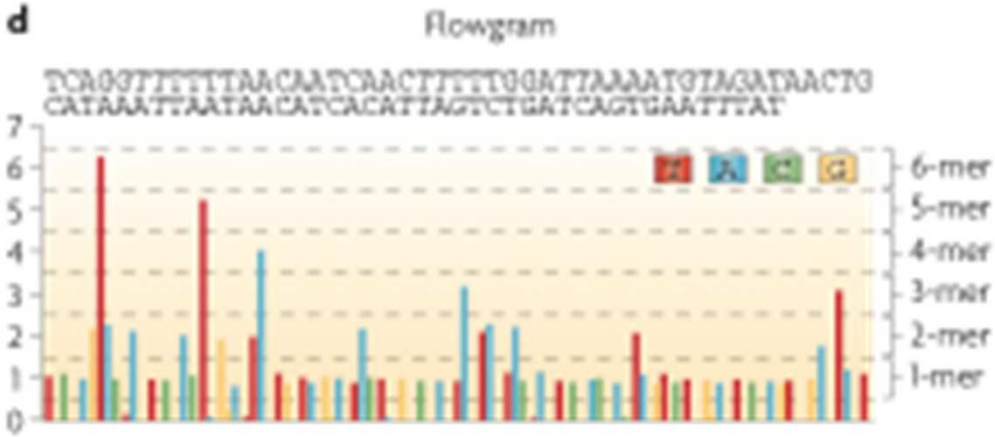
Image issue de Metzker, Nature Rev Genet, 2010

2e génération des méthodes de séquençage

- Principe
  - Préparation de l'échantillon
    - Immobilisation des templates
  - Séquençage et visualisation
    - Cyclic reversible termination (CRT)
    - Sequencing by ligation (SBL)
    - Single-nucleotide addition (SNA)
    - Real-time sequencing

Single-nucleotide addition (SNA) ou pyroséquençage





Images issues de Metzker, Nature Rev Genet, 2010

2e génération des méthodes de séquençage

- Principe
  - Préparation de l'échantillon
    - Immobilisation des templates
  - Séquençage et visualisation
    - Cyclic reversible termination (CRT)
    - Sequencing by ligation (SBL)
    - Single-nucleotide addition (SNA)
    - Real-time sequencing

Real-time sequencing

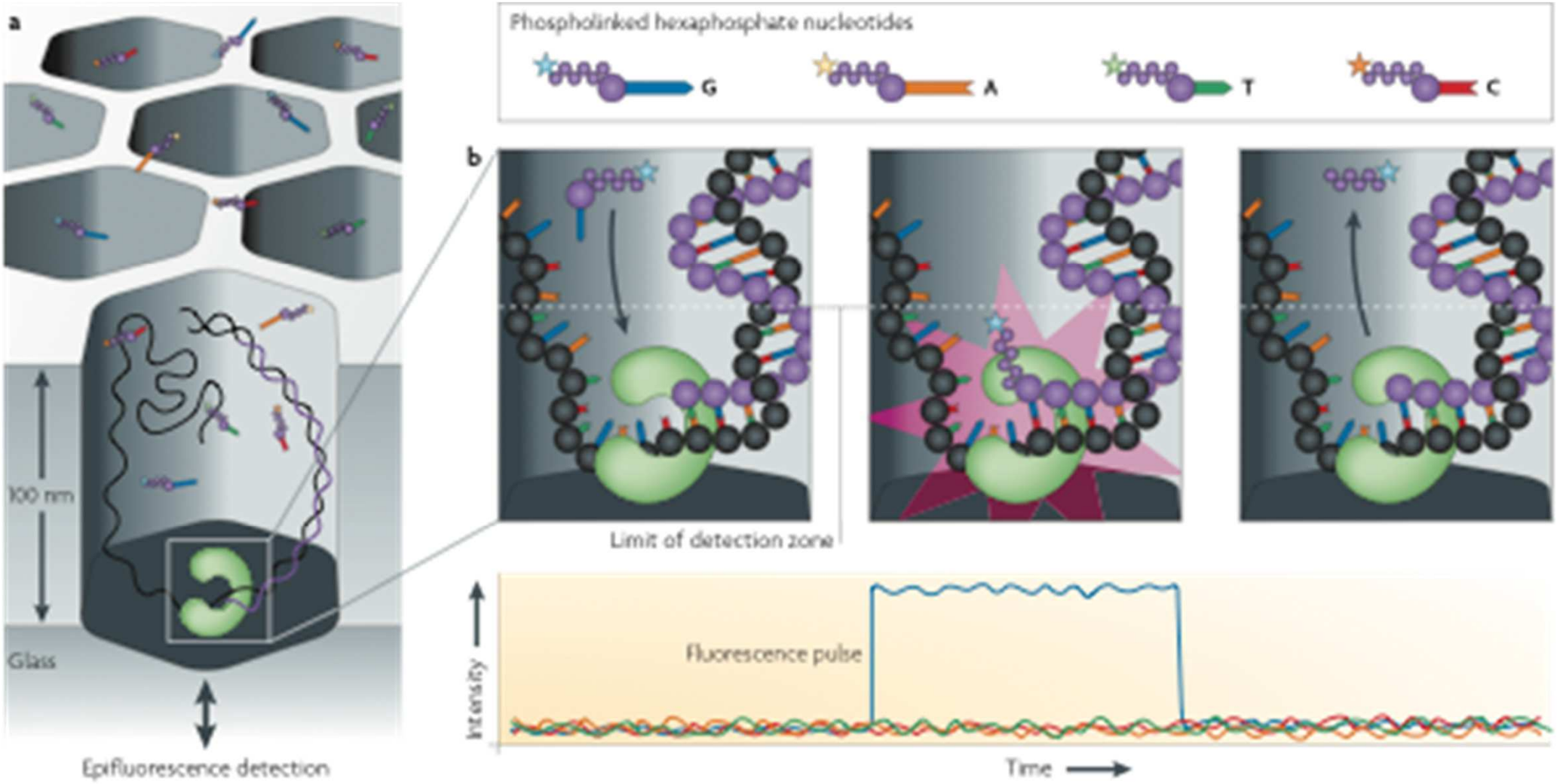


Image issue de Metzker, Nature Rev Genet, 2010

Les différentes plateformes 2e génération

Plateforme	Préparation de l'échantillon	Séquençage
Roche/454	Amplification/PCR par émulsion	Pyroséquençage
Illumina	Amplification/Solid-phase amplification	CRT
Life/APG	Amplification/PCR par émulsion	SBL
Polonator	Amplification/PCR par émulsion	SBL
Helicos BioSciences	Immobilisation des templates ou primers	CRT
Pacific Biosciences	Immobilisation des polymérases	Reat-time sequencing

Tableau inspiré du Tableau 1 de Metzker, Nature Rev Genet, 2010

Avantages

- Haut débit
- Coûts faibles

2e génération des méthodes de séquençage

- Limites
  - Amplification
    - Risque d’erreurs d’amplification
    - Augmentation de la complexité et du temps associé à la préparation
  - Génération importante de données
    - Traitement
    - Stockage
  - Séquences courtes

Entre 2e et 3e génération des méthodes de séquençage

- Ion Torrent
  - Mesure du largage d'un hydrogène lors de l'incorporation d'une base
  - Avantages
    - Pas de besoin de lumière, scanning et caméras pour diriger le processus
  - Inconvénients
    - Toujours un système « wash and scan » avec une amplification PCR

Entre 2e et 3e génération des méthodes de séquençage

- Helicos Genetic Analysis Platform
  - Imagerie d'ADN individuels fixés sur une surface plane alors qu'ils sont alongés en utilisant un primer défini et une polymérase modifiée ainsi que des analogues nucléotides fluorescents
  - Avantages
    - Pas de PCR
  - Inconvénients
    - Arrêt de l'élongation
    - Taux d'erreur > 5%

Séquençage

Méthodes de séquençage

- 1e génération
- 2e génération
- 3e génération
- Comparaison des générations et plateformes

Traitement des données issus du séquençage

- Prétraitements des données
- Assemblage
- Analyse des séquences compilées

3e génération des méthodes de séquençage

3 principales méthodes

- SBS
- Nanopores
- Imagerie directe

3e génération des méthodes de séquençage

3 méthodes

- SBS
- Nanopores
- Imagerie directe

SBS

Observation de molécules d'ADN polymérases simple lors de la synthèse de l'ADN

2 techniques

- Séquençage en temps réel de molécules simples
  - Pacific Biosciences
- Séquençage en temps réel avec observation du transfert d'énergie entre molécule fluorescentes

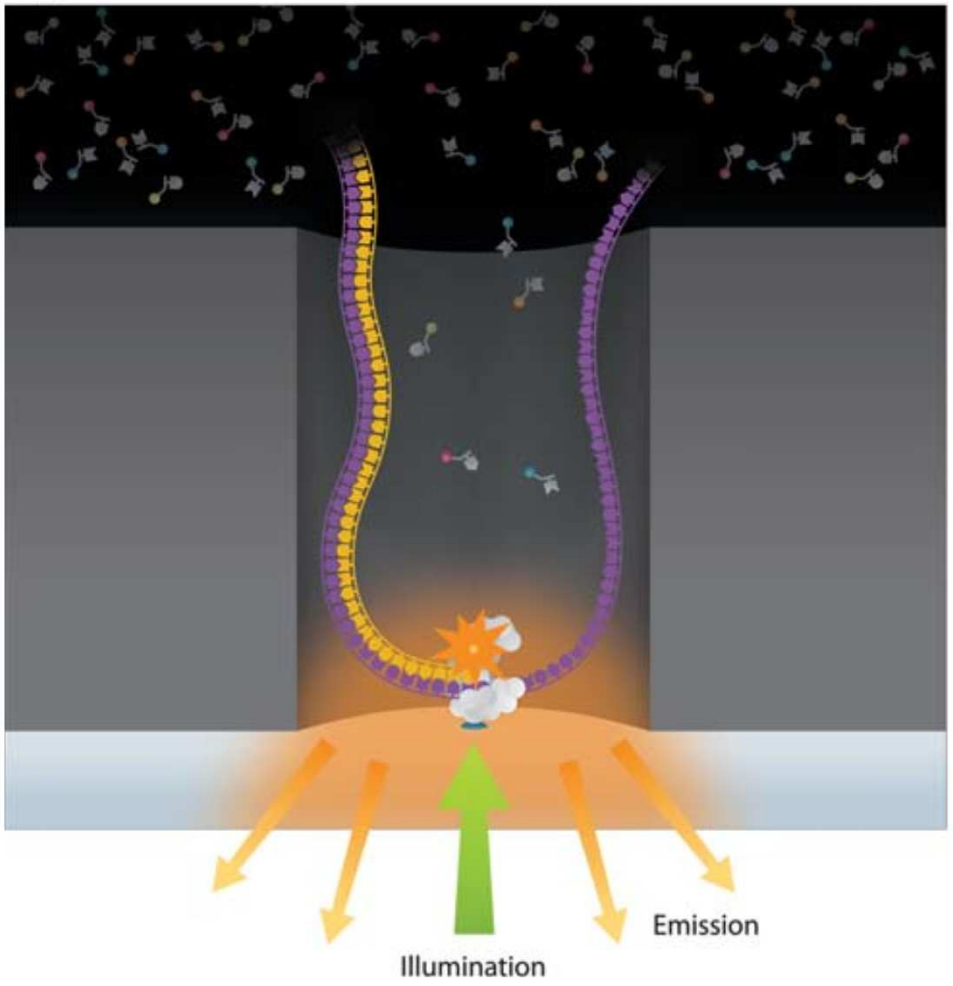


Image issue de Schadt et al, Hum Mol Genet, 2010

3e génération des méthodes de séquençage

3 méthodes

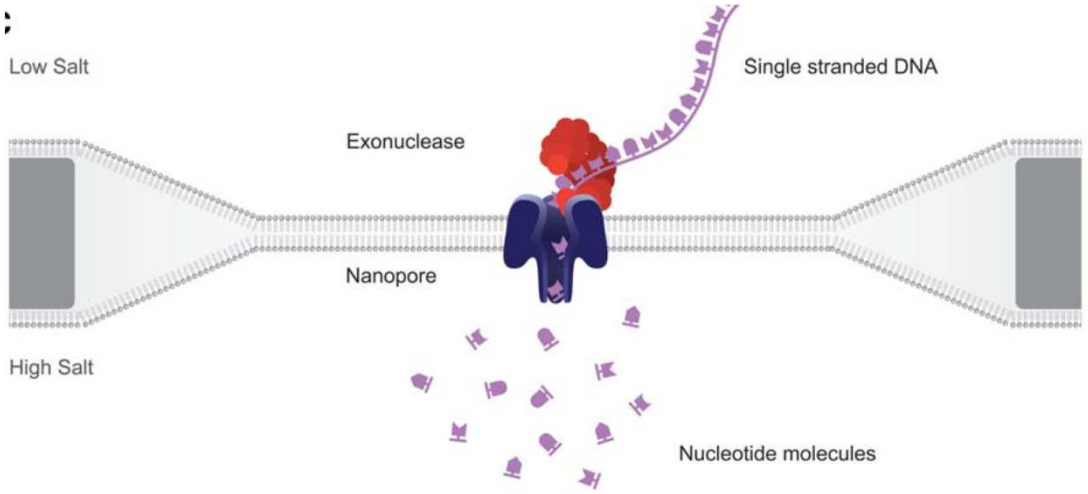
- SBS
- Nanopores
- Imagerie directe

Nanopores

Principe

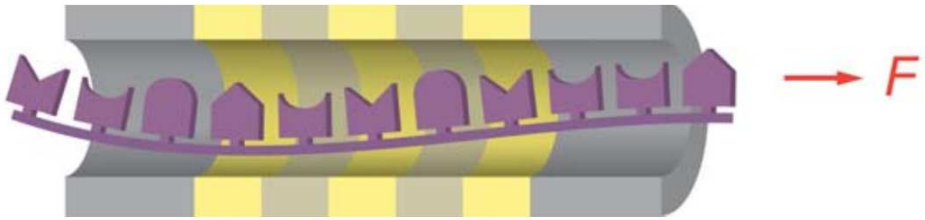
- Enfilage ou positionnement de molécules d’ADN simples à travers ou à proximité d’un nanopore
- Détection des bases individuelles quand elles passent à travers le nanopore

Nanopores



4 méthodes

- Détection électrique directe des molécules d'ADN simples
  - Oxford Nanopore
- Utilisation d'un nanopore biologique comme MspA (*Mycobacterium smegmatis* Porin A)
- Mesure optique
- Séquençage médié par un transistor



Images issues de Schadt et al, Hum Mol Genet, 2010

3e génération des méthodes de séquençage

3 méthodes

- SBS
- Nanopores
- Imagerie directe

Imagerie directe des molécules d'ADN individuelles par microscopie

2 techniques de microscopie pour l'imagerie directe

- Microscopie par émission d'électrons
  - Halcyon Molecular
  - ZS Genetics
- Microscopie à effet tunnel
  - Reveo

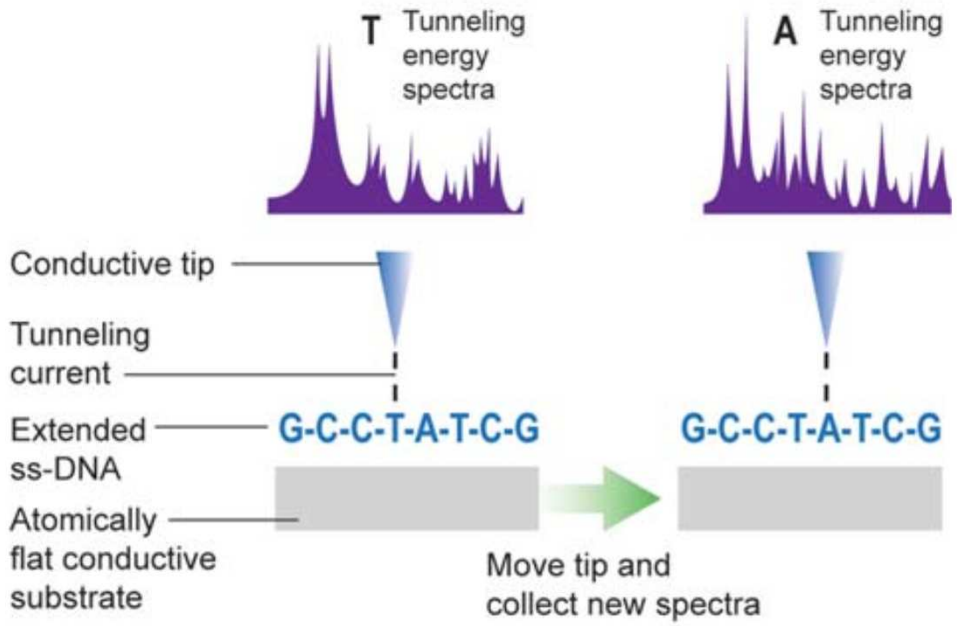


Image issue de Schadt et al, Hum Mol Genet, 2010

3e génération des méthodes de séquençage

- Avantages
  - Molécule simple
  - Haut débit
  - Délai de production plus courts
  - Reads plus longs
    - Facilitation de l'assemblage de novo
  - Plus forte précision
  - Quantité d'ADN nécessaire plus faible
  - Coûts plus faibles

Séquençage

Méthodes de séquençage

- 1e génération
- 2e génération
- 3e génération
- Comparaison des générations et plateformes

Traitement des données issus du séquençage

- Prétraitements des données
- Assemblage
- Analyse des séquences compilées

Comparaison des différentes générations

	1e génération	2e génération	3e génération
Technologie fondamentale	Discrimination par taille de brin, fragment ADN labellisé, produit par SPS ou dégradation	"Wash-and-scan" SPS	SPS, par dégradation, par inspection directe de molécule d'ADN
Résolution	Moyenne en fonction du nombre de copies de l'ADN séquencé	Moyenne en fonction du nombre de copies de l'ADN séquencé	Élevé (une molécule)
Précision	Haute	Haute	Modérée
Taille des séquences lues	800/1000 pb	10/50 pb	1000+ pb
Volume de données en sortie	Bas	Haut	Modéré
Coût	Elevé par base Bas par séquençage	Bas par base Élevé par séquençage	Bas/Modéré par base Bas par séquençage
Temps de séquençage	~ Heures	~ Jours	~ Heures
Préparation des échantillons	Modérement complexe, PCR non requise	Complexe, PCR requise	Simple/Complexe suivant la technologie employée
Analyse des données	Facile	Complexe à cause du nombre important des données et de la petite taille des reads qui compliquent l'assemblage des génomes et des algorithmes d'alignement	Complexe : nombre important des données, ces nouvelles technologies émettent de nouveaux types d'informations et donc de nouvelles doivent être implémentées
Résultats primaires	Identification des bases avec des scores de qualité	Identification des bases avec des scores de qualité	Identification des bases avec des scores de qualité, et d'autres informations comme la cinétique, la modification des bases

Tableau inspiré du Tableau 1 de Schadt et al, Hum Mol Genet, 2010

Quelle plateforme choisir?

À prendre en considération

- Application souhaitée
- Coût d’un run et de son traitement
- Longueur des reads
- Nombre de reads par run
- Taux d’erreur
- Disponibilité

Longueur des reads

Longueur des reads (bp)	Éléments du génome
25 - 75	SNP, short frameshift mutations
100 - 400	Short functional signatures
500 - 1 000	Whole domain, single domain genes
1 000 - 5 000	Short operons, multidomain genes
5 000 - 10 000	Longer operons, some cis-control elements
> 100 000	Prophages, pathogenicity islands, various mobile insertion elements
> 1 000 000	Whole prokaryotic chromosome organization

Tableau inspiré du Tableau 2 de Wooley et al, Plos Comp Biol, 2010

- Reads longs
  - Facilité d'assemblage
  - Optimal pour des génomes jamais séquencés et la caractérisation du transcriptome
- Reads courts
  - Faibles coûts
  - Plus forte couverture
  - Re-séquençage pour applications basées sur le fréquence (comptage)

Instrument	Run time	Millions of Reads/run	Bases / read	Reagent Cost/run	Reagent Cost/Gb	Reagent Cost/Mread
Applied Biosystems 3730 (capillary)	2 hrs,	0,000096	650	$144	$2307692,31	$1500000,00
454 GS Jr, Titanium	10 hrs,	0,1	400	$977	$19540,00	$9770,00
454 FLX Titanium	10 hrs,	1	400	$6200	$15500,00	$6200,00
454 FLX+	20 hrs,	1	650	$6200	$9538,46	$6200,00
Illumina GA IIx - v5 SE	2 days	640	36	$4842	$210,16	$7,57
Illumina GA IIx - v5 PE	14 days	640	288	$17978	$97,54	$28,09
Illumina MiSeq v2 Nano	17 hrs,	1	300	$530	$1766,67	$530,00
Illumina MiSeq v2 Nano	28 hrs,	1	500	$639	$1278,00	$639,00
Illumina MiSeq v2 Micro	19 hrs,	4	300	$798	$665,00	$199,50
Illumina MiSeq v2	5 hrs,	15	50	$747	$996,00	$49,80
Illumina MiSeq v2	24 hrs,	15	300	$958	$212,89	$63,87
Illumina MiSeq v2	39 hrs,	15	500	$1066	$142,13	$71,07
Illumina MiSeq v3	20 hrs,	22	150	$824	$249,70	$37,45
Illumina MiSeq v3	55 hrs,	22	600	$1442	$109,24	$65,55
Illumina NextSeq 500	15 hrs,	130	150	$975	$50,00	$7,50
Illumina NextSeq 500	26 hrs,	130	300	$1560	$40,00	$12,00
Illumina NextSeq 500	11 hrs,	400	75	$1300	$43,33	$3,25
Illumina NextSeq 500	18 hrs,	400	150	$2500	$41,67	$6,25
Illumina NextSeq 500	30 hrs,	400	300	$4000	$33,33	$10,00
Illumina HiSeq 2500 - rapid run	10 hrs,	300	50	$1350	$90,00	$4,50
Illumina HiSeq 2500 - rapid run	27 hrs,	300	200	$3126	$52,10	$10,42
Illumina HiSeq 2500 - rapid run	40 hrs,	300	300	$4126	$45,84	$13,75
Illumina HiSeq 2500 - high output v3	2 days	1500	50	$5866	$78,21	$3,91
Illumina HiSeq 2500 - high output v3	11 days	1500	200	$13580	$45,27	$9,05
Illumina HiSeq 2500 - high output v4	40 hrs,	2000	50	$5866	$58,66	$2,93
Illumina HiSeq 2500 - high output v4	6 days	2000	250	$14950	$29,90	$7,48
Illumina HiSeq X (2 flow cells)	3 days	6000	300	$12750	$7,08	$2,13
Ion Torrent – PGM 314 chip	2,3 hrs,	0,475	200	$349	$3673,68	$734,74
Ion Torrent – PGM 314 chip	3,7 hrs,	0,475	400	$474	$2494,74	$997,89
Ion Torrent – PGM 316 chip	3 hrs,	2,5	200	$549	$1098,00	$219,60
Ion Torrent – PGM 316 chip	4,9 hrs,	2,5	400	$674	$674,00	$269,60
Ion Torrent – PGM 318 chip	4,4 hrs,	4,75	200	$749	$788,42	$157,68
Ion Torrent – PGM 318 chip	7,3 hrs,	4,75	400	$874	$460,00	$184,00
Ion Torrent - Proton I	4 hrs,	70	175	$1000	$81,63	$14,29
Ion Torrent - Proton II (forecast)	5 hrs,	280	175	$1000	$20,41	$3,57
Ion Torrent - Proton III (forecast)	6 hrs,	500	175	$1000	$11,43	$2,00
Life Technologies SOLiD – 5500xl	8 days	1410	110	$10503	$67,72	$7,45
Pacific Biosciences RS II	2 hrs,	0,03	3000	$100	$1111,11	$3333,33
Oxford Nanopore MinION (forecast)	≤6 hrs,	0,1	9000	$900	$1000,00	$9000,00
Oxford Nanopore GridION 2000 (forecast)	varies	4	10000	$1500	$37,50	$375,00
Oxford Nanopore GridION 8000 (forecast)	varies	10	10000	$1000	$10,00	$100,00

Source : http://www.molecularecologist.com/next-gen-fieldguide-2014/

Multiplexage

Moyen de diminuer les coûts de séquençage

- Ajout d’un tag d’identification pour chaque échantillon
- Mélange des échantillons
- Préparation et séquençage en parallèle
- Tri des séquences en fonction des échantillons, l’information de source étant contenue dans la séquence

Taux d’erreurs

Instrument	**Erreurs principales**	**Taux d'erreur final (%)**
3730xl (capillary)	Substitutions	0.1-1
454 (tous modèles)	Indels	1
Illumina (tous modèles)	Substitutions	0.1
Ion Torrent (tous modèles)	Indels	1
SOLiD - 5500 xl	Biais AT	<=0.1
Oxford Nanopore	Délétions	4
PacBio RS	Indels	<=1

Source : http://www.molecularecologist.com/next-gen-fieldguide-2014/

Comparaison des plateformes

Instrument	Principaux avantages	Principaux désavantages
3730xl (capillary)	Faibles coûts pour des très petites études	Prix très élevés pour des grandes quantités de données
Illumina MiSeq	Coûts modérés d'instruments et runs; Faibles coûts par Mb pour une petite plateforme; Durée des runs les plus rapides et reads les longs des plateformes Illumina	Relativement peu de reads et coûts plus élevés par Mb par rapport à HiSeq
Illumina NextSeq 500	Facilité d'utilisation; Coûts modérés d'instruments et de run;...	Nouvelle chimie et nouveau design d'instruments
Illumina HiSeq 2500	Faibles coûts par Mb; ...	Coûts élevés d'instruments et de runs; Besoin de personnels entrainés; Pas de possibilité d'avoir en même temps un run rapide et une sortie flow cell rapide
Illumina HiSeq X	Plus forte capacité et plus faible coût par base que HiSeq 2500	Pas disponible avec 2015
Illumina HiSeq X Ten	Plus faibles coûts par Mb	Seulement 2 systèmes disponibles en 2014; Besoin de 10M$ de stockage de données
Ion Torrent - PGM	Faible coût d'instruments; Machine très simple; ...	Plus fort taux d'erreur qu'Illumina; Plus de temps de manipulation et moins de reads à coûts plus élevés que MiSeq; Communauté d'utilisation plus petites; ...
Ion Torrent - Proton	Coûts modéremment faible d'instrument pour des applications haut débit; ...	Plus forts taux d'erreur, plus de temps de manipulation and plus petites bases de données globale qu'Illumina; Coûts plus élevés par Mb que la plupart des instruments Illumina
PacBio	Séquençage en temps réel simple molécule; Plus long reads disponibles; Capacité de détecter des modifications de bases; ...	Forts taux d'erreurs; Faible nombre de reads par run; Coûts élevés par Mb, ...
454 GS Jr.	Plus faible coûts par run que 454 FLX+	Coûts élevés par Mb; Peu de reads; Reads plus courts que FLX
454 FLX+	Reads contigus les plus longs pour du 2e génération	Coûts élevés par Mb; Problèmes de reagent; ...
SOLiD - 5500/5500xl/5500W	Forte précision; Capacité de sauver les cycles de séquençage ayant échoué	Longévité de la plateforme; Reads relativement courts; Plus de gaps dans l'assemblage que pour les données Illumina; ...
Oxford Nanopore (projet)	Instrument portable et low-cost; Reads extrement longs possibles; ...	Instruments par encore disponibles; Pas de données publiquement disponibles; Informations disponibles limitées; ...
Oxford Nanopore GridIon (projet)	Reads extrement longs possibles; Instrument low-cost; Pas d'augmentation des taux d'erreur avec la longueur des reads; ...	Pas disponible; Pas de données publiquement disponibles;4% de taux d'erreur; ...

Source : http://www.molecularecologist.com/next-gen-fieldguide-2014/

Séquençage

Méthodes de séquençage

- 1e génération
- 2e génération
- 3e génération
- Comparaison des générations et plateformes

Traitement des données issus du séquençage

- Prétraitements des données
- Assemblage
- Analyse des séquences compilées

Séquençage

Méthodes de séquençage

- 1e génération
- 2e génération
- 3e génération
- Comparaison des générations et plateformes

Traitement des données issus du séquençage

- Prétraitements des données
- Assemblage
- Analyse des séquences compilées

Prétraitements des données

- Gestion des données
  - Formats
  - Types de données
- Assignation des reads multiplexés aux échantillons et suppression des primers et adaptateurs
- Filtrage des reads en fonction de la qualité
- Réduction des biais de séquençage

Prétraitements des données

- Gestion des données
  - Formats
  - Types de données
- Assignation des reads multiplexés aux échantillons et suppression des primers et adaptateurs
- Filtrage des reads en fonction de la qualité
- Réduction des biais de séquençage

Différents formats de fichiers

- FastA
  - Format répandu de stockage des séquences biologiques
- FastQ
  - Illumina et d'autres séquenceurs
- SFF
  - Roche 454
- SRF
  - Helicos
- HDF5
  - PacBio, Applied Biosystems, Oxford Nanopore

Les plus utilisés

Différents formats de fichiers

- FastA
  - Format répandu de stockage des séquences biologiques
- FastQ
  - Illumina et d'autres séquenceurs
- SFF
  - Roche 454
- SRF
  - Helicos
- HDF5
  - PacBio, Applied Biosystems, Oxford Nanopore

FastA

Base de données	Format de l'identification
GenBank	*gi*|*numéro gi*|*gb*|*numéro d'accession*|*locus*
EMBL	*gi*|*numéro gi*|*emb*|*numéro d'accession*|*locus*
DNA Data Bank of Japan	*gi*|*numéro gi*|*dbj*|*numéro d'accession*|*locus*
NBRF PIR	*pir*||*entrée*
Protein Research Foundation	*prf*||*nom*
Swiss-Prot	*sp*|*numéro d'accession*|*nom*
Brookhaven Protein Data Bank	*pdb*|*entrée*|*chaine*
Brevets	*pat*|*brevet*|*numéro*
GenInfo Backbone Id	*bbs*|*numéro*
General database identifier	*gnl*|*base de données*|*identifiant*
NCBI Reference Sequence	*ref*|*numéro d'accession*|*locus*
Local Sequence identifier	*lcl*|*identifiant*

Identifiant unique
Pas de description définitive mais tentative

```
>Identifiant (Commentaire)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
>Identifiant2 (Commentaire)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXX
```

Suite de lettres représentant les acides nucléiques ou les acides aminés
Lignes de 120 résidus maximum

Code des acides nucléiques

Code des acides nucléiques	**Signification**	**Moyen de mémorisation mnémotechnique**
A	A	**A**dénosine
C	C	**C**ytosine
G	G	**G**uanine
T	T	**T**hymine
U	U	**U**racile
R	A ou G	pu**R**ine
Y	C, T ou U	p**Y**rimidine
K	G, T ou U	bases contenant une cétone (**K**étones en anglais)
M	A ou C	bases contenant un groupe a**M**ine
S	C ou G	interaction forte (**S**trong en anglais)
W	A, T ou U	interaction faible (**W**eak en anglais)
B	différent de A (C, G, T ou U)	**B** vient après A
D	différent de C (A, G, T ou U)	**D** vient après C
H	différent de G (A, C, T ou U)	**H** vient après G
V	différent de T et U (A, C ou G)	**V** vient après U
N	A, C, G, T ou U	**N**'importe ou **N**ucléotide
X	acide nucléique masqué	
-	gap	

Code des acides aminés
Standard IUB/IUPAC

Code des acides aminés	Signification
A	Alanine
B	Acide aspartique ou Asparagine
C	Cystéine
D	Acide aspartique
E	Acide glutamique
F	Phénylalanine
G	Glycine
H	Histidine
I	Isoleucine
K	Lysine
L	Leucine
M	Méthionine
N	Asparagine

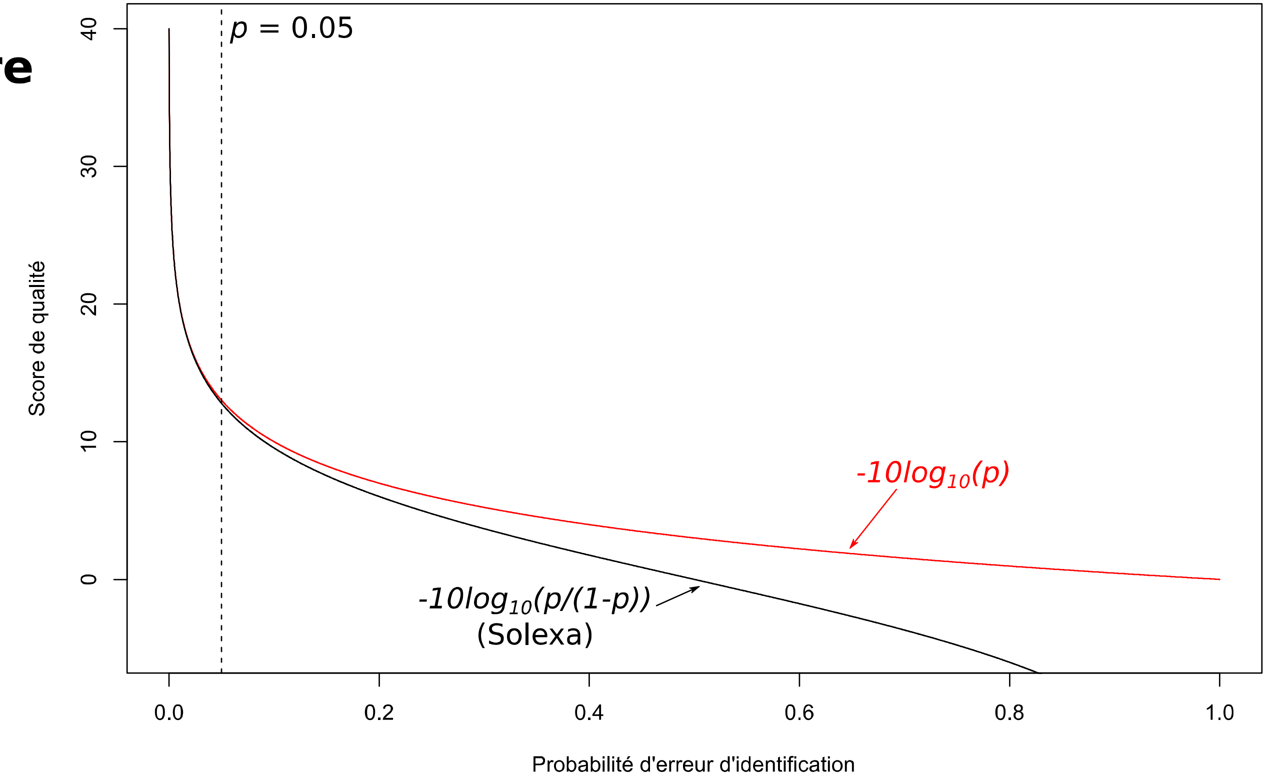
Code des acides aminés	Signification
O	Pyrrolisine
P	Proline
Q	Glutamine
R	Arginine
S	Sérine
T	Thréonine
U	Sélénocystéine
V	Valine
W	Tryptophane
Y	Tyrosine
Z	Acide glutamique ou Glutamine
X	N'importe
*	Codon stop
-	gap

Différents fichiers au format FastA

Extension	Signification	Commentaires
.fasta, *.fas*, *.fa*	fasta générique	Tout fichier fasta. Ces types de fichiers peuvent avoir aussi comme extension *.seq* (pour séquence) et *.fsa* (pour fasta sequence alignment)
.fna	fasta nucleic acid	Fichier fasta contenant une séquence d'acides nucléiques. Pour des séquences de régions codants spécifiques d'un génome, il sera préféré l'extension *.ffn*
.ffn	fasta functional nucleotide	Fichier fasta contenant une séquence nucléique d'une région codante d'un génome
.faa	fasta amino acid	Fichier fasta contenant une séquence d'acides aminés. Un fichier contenant de multiples séquences pourra avoir l'extension plus spécifique *mpfa*
.frn	fasta RNA non-coding	Fichier fasta contenant une séquence d'ARN non-codant d'un génome (comme ARNt et ARNr) mais écrite dans la nomenclature du code ADN

Qualité des séquences Fasta
Fichier QUAL avec des scores PHRED

Calcul du score de qualité



Codage du score de qualité dans un fichier à part

```
>Identifiant (Commentaire)
X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X
X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X X
X X X X X X X X X X X X X X X X X X X X X X X X X X
>Identifiant2 (Commentaire)
```

Suite de chiffres correspondant au score de qualité pour chacune des bases de la séquence

Différents formats de fichiers

- FastA
  - Format répandu de stockage des séquences biologiques
- FastQ
  - Illumina et d'autres séquenceurs
- SFF
  - Roche 454
- SRF
  - Helicos
- HDF5
  - PacBio, Applied Biosystems, Oxford Nanopore

FastQ

Identifiant unique comme après le ">" en fasta

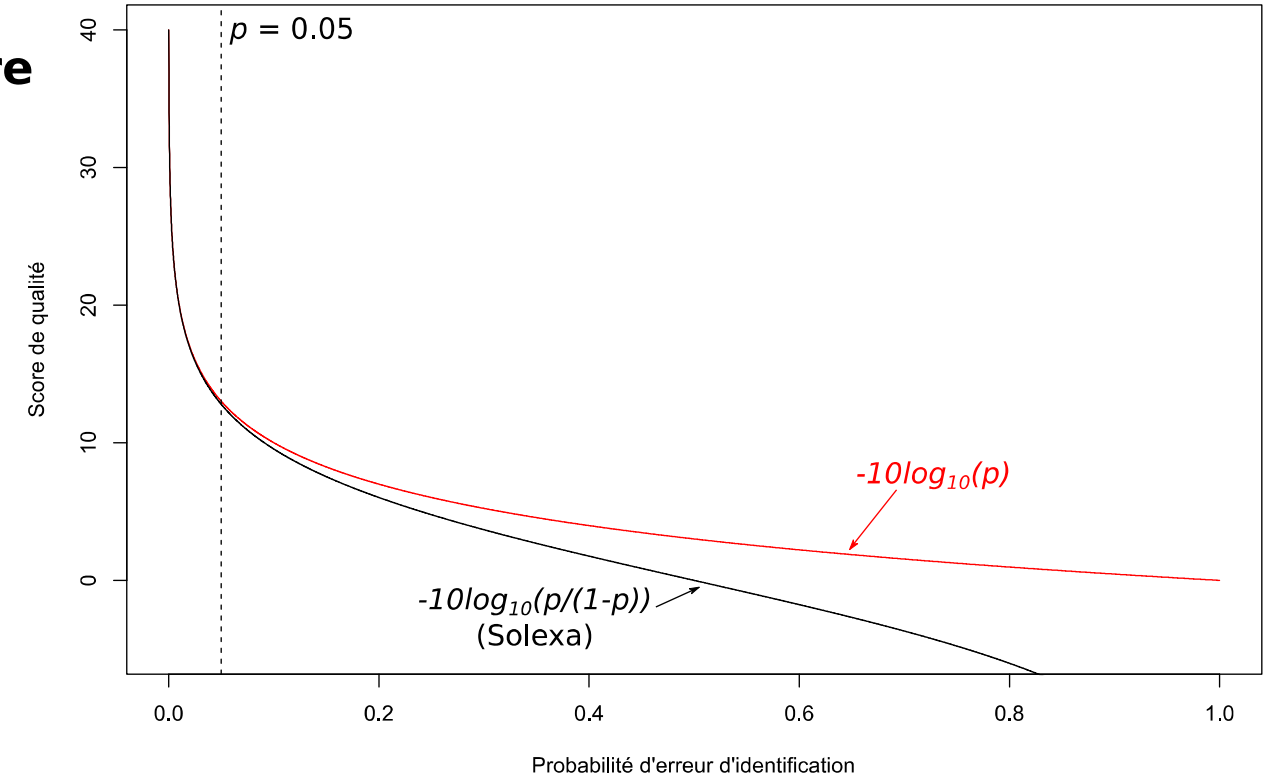
Séquence nucléique brute

```
@Identifiant
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX...
+
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX...
```

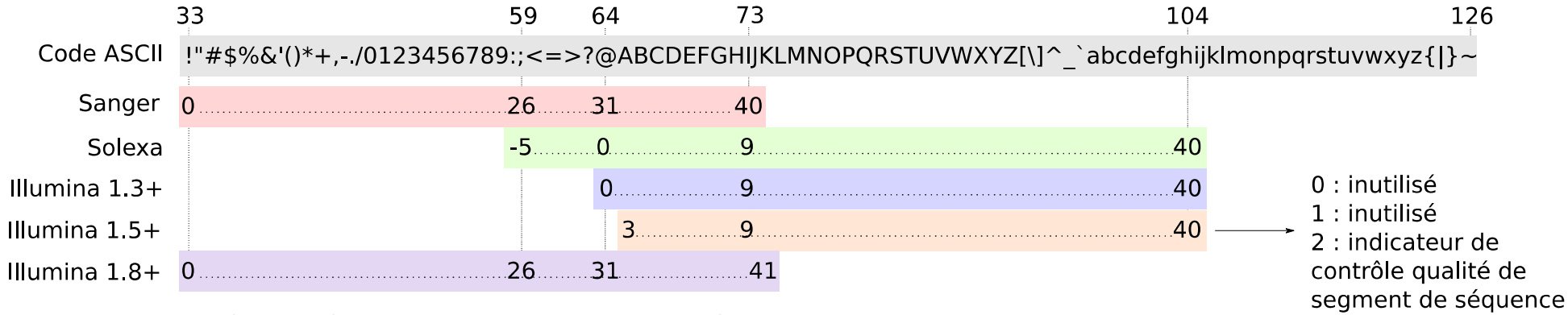
Scores de qualité associés à chacune des bases de la séquence : suite de caractères ASCII entre le 33e et le 127e

Score de qualité dans FastQ

Calcul du score de qualité



Codage du score de qualité



Différents formats de fichiers

- FastA
  - Format répandu de stockage des séquences biologiques
- FastQ
  - Illumina et d'autres séquenceurs
- SFF
  - Roche 454
- SRF
  - Helicos
- HDF5
  - PacBio, Applied Biosystems, Oxford Nanopore

Standard Flowgram Format

Format de sortie de 454 et Ion Torrent

Fichier binaire contenant

- Section en-tête commune à tous les reads du fichier
- Pour chaque read
  - Section en-tête du read
  - Section des données du read

En-têtes d’un fichier SFF

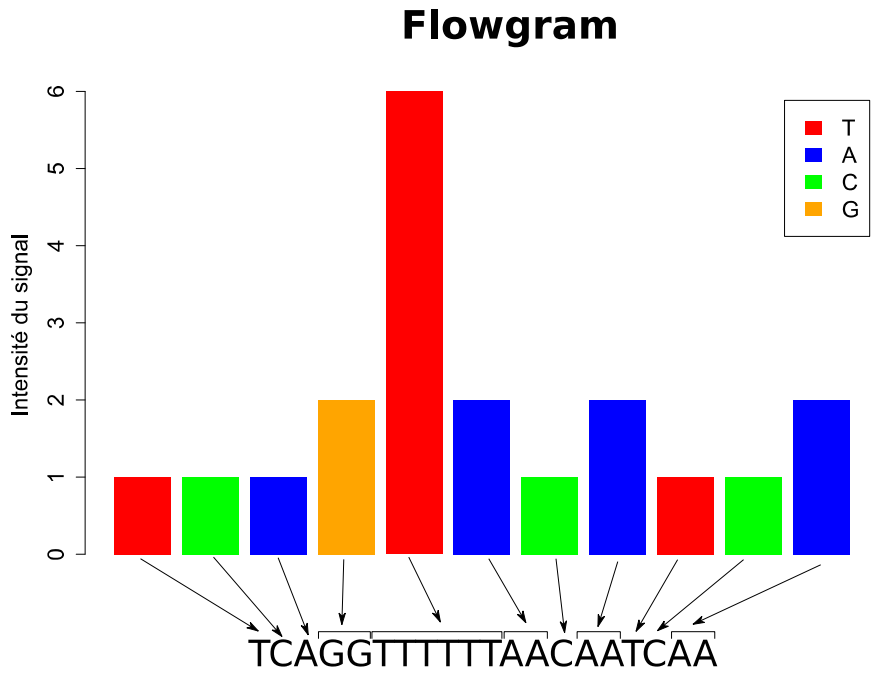
En-tête globale d'un fichier

magic_number	uint32_t	
version	char[4]	
index_offset	uint64_t	*Offset et longueur d'un index optionel des reads*
index_length	uint32_t	
number_of_reads	uint32_t	*Nombre de reads contenus dans le fichier*
key_length	uint16_t	*Longueur et bases nucléotides de la séquence clé utilisée pour ces reads*
key_sequence	char[key_length]	
header_length	uint16_t	*Nombre d'octets requis pour l'ensemble des champs de l'en-tête*
number_of_flows_per_read	uint16_t	*Nombre de flow pour chaque read*
flowgram_format_code	uint8_t	*Format utilisé pour encoder chaque valeur du flowgram de chaque read*
flow_chars	char[number_of_flows_per_read]	*Tableau des bases nucléotides correspondant aux nucléotides utilisés pour chaque flow de chaque read*
eight_byte_padding	uint8_t[*]	
index_magic_number	uint32_t	*S'il y a un index des reads*
Index_version	char[4]	

En-tête d'un read

read_header_length	uint16_t	*Longueur de l'en-tête du read en octet*
name_length	uint16_t	*Longueur et nom de l'accession ou nom du read*
name	char[name_length]	
number_of_bases	uint32_t	*Nombre de bases dans ce read*
clip_qual_left	uint16_t	*Position et qualité de la première base après le point de coupure au début du read*
clip_adapter_left	uint16_t	
clip_qual_right	uint16_t	*Position et qualité de la dernière base avant le point de coupure à la fin du read*
clip_adapter_right	uint16_t	
eight_byte_padding	uint8_t[*]	

Données de séquence dans un fichier SFF



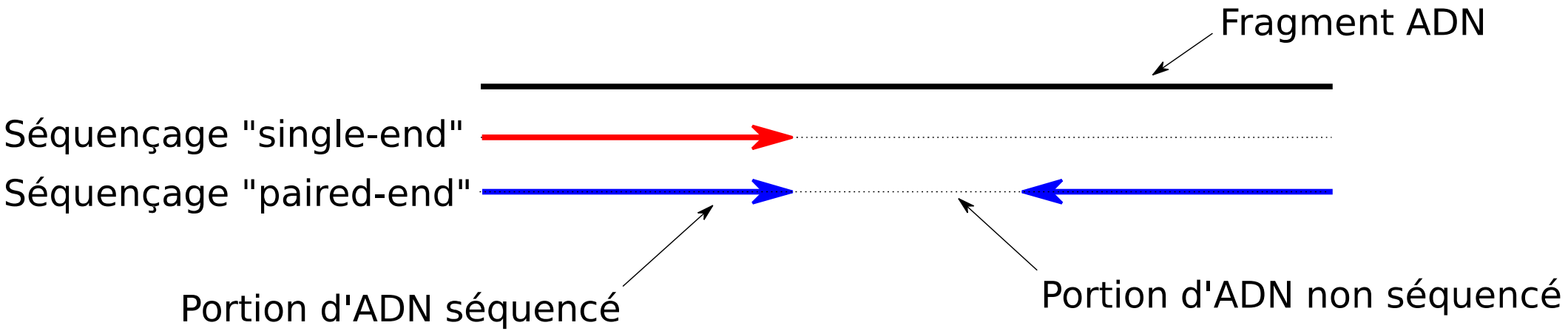
Enregistrement des données de séquence d'un read

flowgram_values	uint*_t[number_of_flows]	*Estimation de l'étirement de l'homopolymère pour chaque flow*
flow_index_per_base	uint8_t[number_of_bases]	*Position de flow de chaque base dans la séquence*
bases	char[number_of_bases]	*Séquence de nucléotides*
quality_scores	uint8_t[number_of_bases]	*Score de qualité pour chaque base (même calcul que pour les données fastq)*
eight_byte_padding	uint8_t[*]	

Prétraitements des données

- Gestion des données
  - Formats
  - Types de données
- Assignation des reads multiplexés aux échantillons et suppression des primers et adaptateurs
- Filtrage des reads en fonction de la qualité
- Réduction des biais de séquençage

Séquençage « single-end » ou « paired-end »



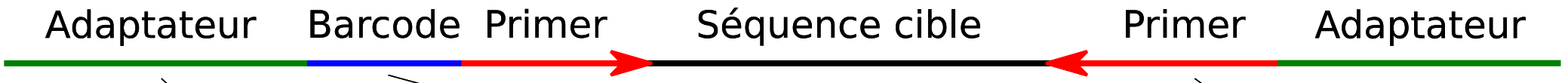
Séquençage « paired-end »

- Génération de deux fichiers de sortie (1 pour chaque bout)
- Concaténation des fichiers en faisant attention au sens de lecture

Prétraitements des données

- Gestion des données
  - Formats
  - Types de données
- Assignation des reads multiplexés aux échantillons et suppression des primers et adaptateurs
- Filtrage des reads en fonction de la qualité
- Réduction des biais de séquençage

Primers, adapters, barcode et démultiplexage



Courte séquence nucléotidique capable de réaliser le pontage entre deux fragments d'ADN terminés par des séquences non complémentaires

Identifiant unique de l'échantillon dans les cas de multiplexage
MID ("Multiplex Identifier", 454) ou Index (Illumina)

Brin d'ADN qui sert de point de départ à la synthèse ADN

Métadonnées de description des échantillons

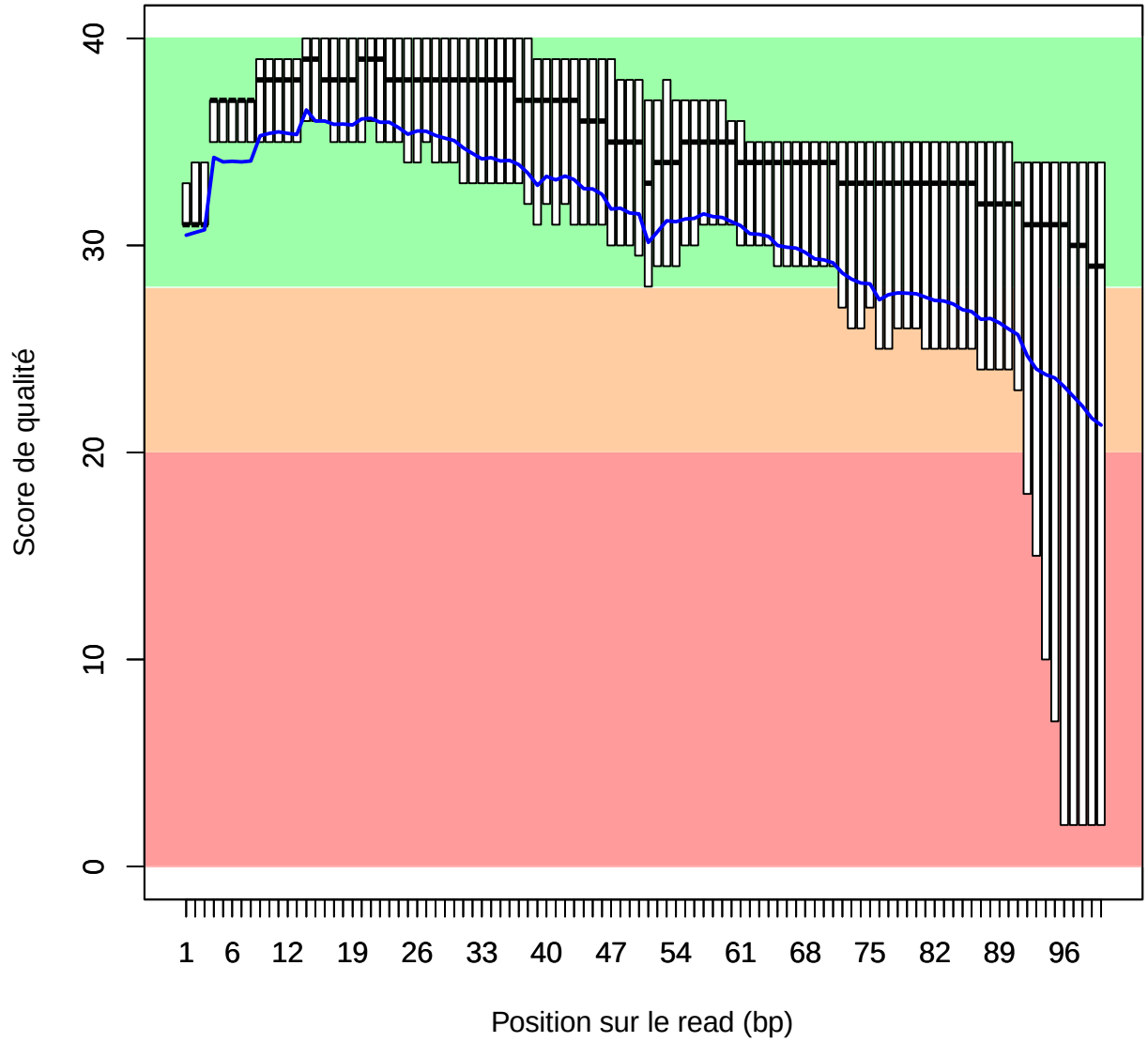
Exemple de "Mapping file"

Sample ID	Barcode	Primer	Reverse Primer	Type	Experiment	...
G1	CGTTTC	ACGGRAGGCAGGCAG	TACCAGGGTATCTAATCCT	DNA	Crohn	...
G2	CCCGTT	ACGGRAGGCAGGCAG	TACCAGGGTATCTAATCCT	DNA	Healthy	...
G3	GGTCAC	ACGGRAGGCAGGCAG	TACCAGGGTATCTAATCCT	RNA	Crohn	...
G4	AGTGCT	ACGGRAGGCAGGCAG	TACCAGGGTATCTAATCCT	RNA	Healthy	...

Prétraitements des données

- Gestion des données
  - Formats
  - Types de données
- Assignation des reads multiplexés aux échantillons et suppression des primers et adaptateurs
- Filtrage des reads en fonction de la qualité
- Réduction des biais de séquençage

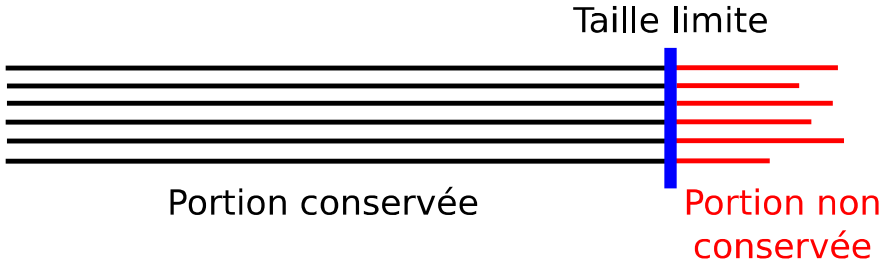
Diminution du score de qualité avec la longueur des reads



Plusieurs solutions de filtrage des reads

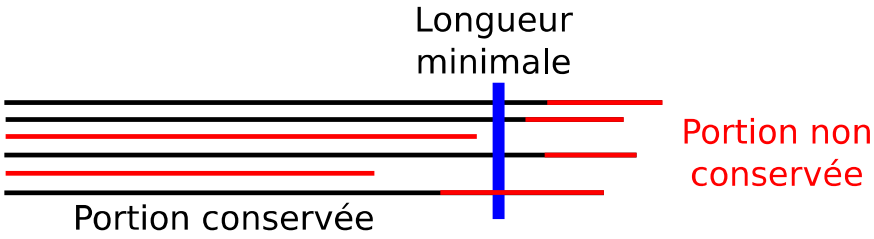
Filtrage de longueur fixe

- Coupure à une longueur donnée



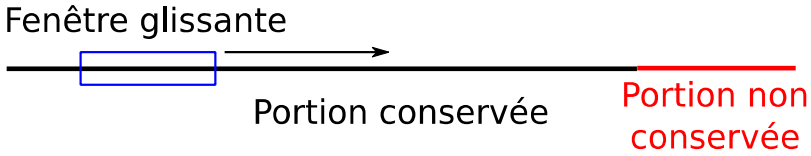
Filtrage adaptatif

- Coupure en fonction du score de qualité
- Longueur minimale de séquence



Filtrage sur une fenêtre glissante

- Coupure en fonction du score de qualité
- Utilisation de la valeur moyenne de la fenêtre



Prétraitements des données

- Gestion des données
  - Formats
  - Types de données
- Assignation des reads multiplexés aux échantillons et suppression des primers et adaptateurs
- Filtrage des reads en fonction de la qualité
- Réduction des biais de séquençage et d’expériences

Séquençage

Méthodes de séquençage

- 1e génération
- 2e génération
- 3e génération
- Comparaison des générations et plateformes

Traitement des données issus du séquençage

- Prétraitements des données
- Assemblage
- Analyse des séquences compilées

Terminologie

Assemblage

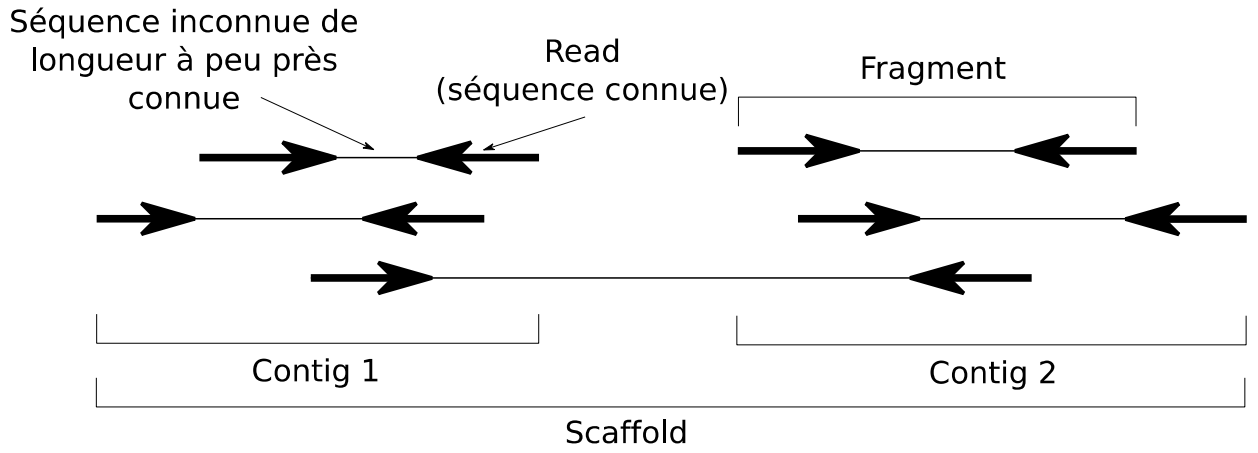
Alignement et fusion de reads en séquence ADN plus longue pour reconstruire la séquence originale

Contig

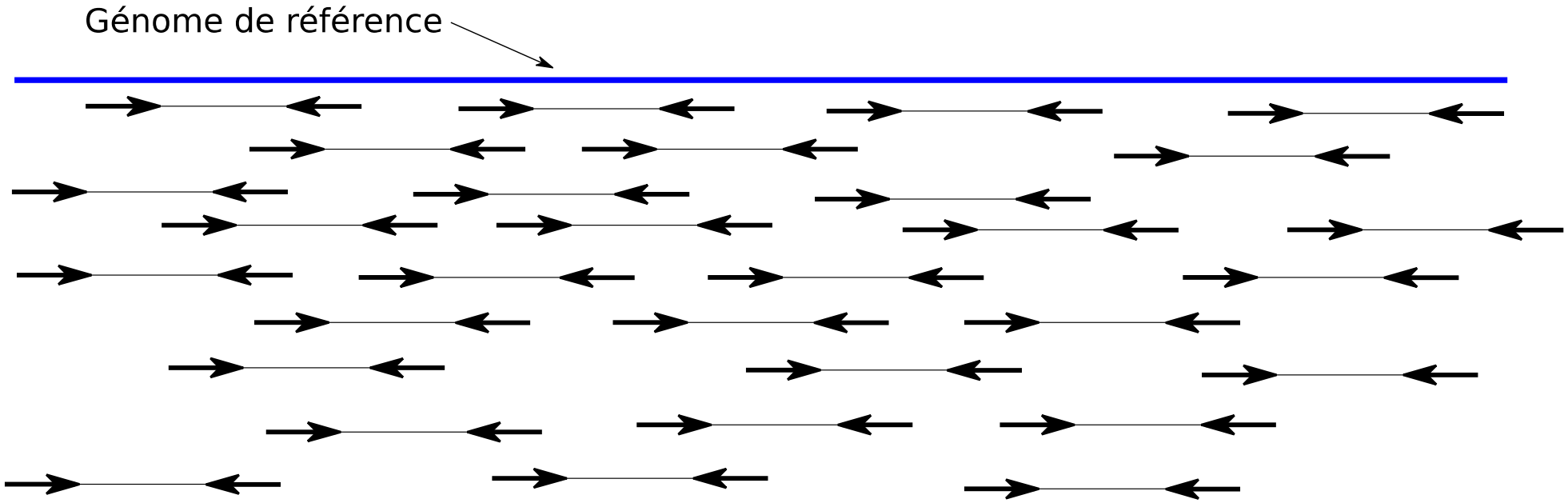
Séquence génomique continue et ordonnée générée par l'assemblage de reads qui se chevauchent

Scaffold

Contigs chevauchant séparés par des gaps de longueur connue



Alignement sur des génomes de référence



Utile pour identifier des variants ou lors du reséquençage de génomes

Assemblage *de novo*

Reconstruction des séquences ADN d'un organismes à partir des seules séquences (pas d'utilisation de génome de référence)

Idéal

- Longs reads sans erreurs
- Problème de simple déduction

Réalité

- Reads courts et sujets aux erreurs
- Problème d'inférence compliqué

Statistiques de séquençage

Profondeur ou couverture

Nombre moyen de fois qu'un nucléotide particulier est représenté dans une collection de reads aléatoires

Profondeur de couverture

Nombre de reads x Longueur des reads / Taille de l'assemblage

Métriques d'assemblage

- Nombre de contigs/scaffolds
- Taille des contigs/scaffold
- Taille totale
- Nombre de « N »
- N50
  - Longueur du plus petit contig dans l'ensemble qui contient le moins de contigs dont les longueurs combinées représente au moins 50% de l'assemblage
  - Exemple
    - Soit 7 contigs de longueur 1,1,3,5,8,12 et 20
    - Longueur totale des contigs = 50
    - Longueur totale des contigs / 2 = 25
    - N50 = 12 car 1+1+3+5+8+12=30 (>25)

Séquençage

Méthodes de séquençage

- 1e génération
- 2e génération
- 3e génération
- Comparaison des générations et plateformes

Traitement des données issus du séquençage

- Prétraitements des données
- Assemblage
- Analyse des séquences compilées

Analyse des séquences compilés

Détection des SNP ou indels

Repérage des variants génétiques

Détection de nouveaux gènes ou d'éléments régulateurs

Détermination des niveaux d'expression

…