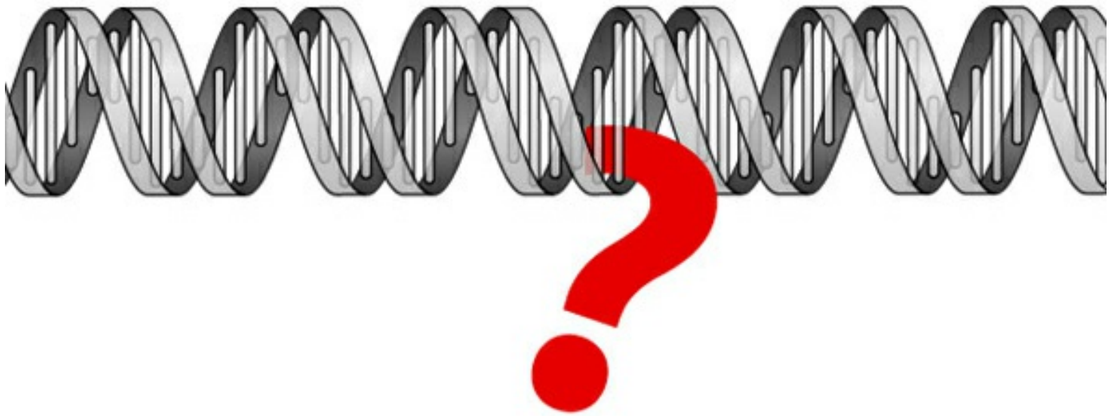
Séquençage

Février 2016

Bérénice Batut

berenice.batut@udamail.fr

Séquençage



Crédits : http://scienceprogress.org

Séquençage ADN

Détermination de l'ordre d'enchaînement des nucléotides d'un fragment d'ADN/ARN donné

Pourquoi séquencer l'ADN?

Identification de gènes, de régions plus larges, de chromosomes ou de génomes entiers

Applications

- Biologie moléculaire
- Biologie évolutive
- Métagénomique
- Médecine
- Forensiques

Un peu de bibliographie

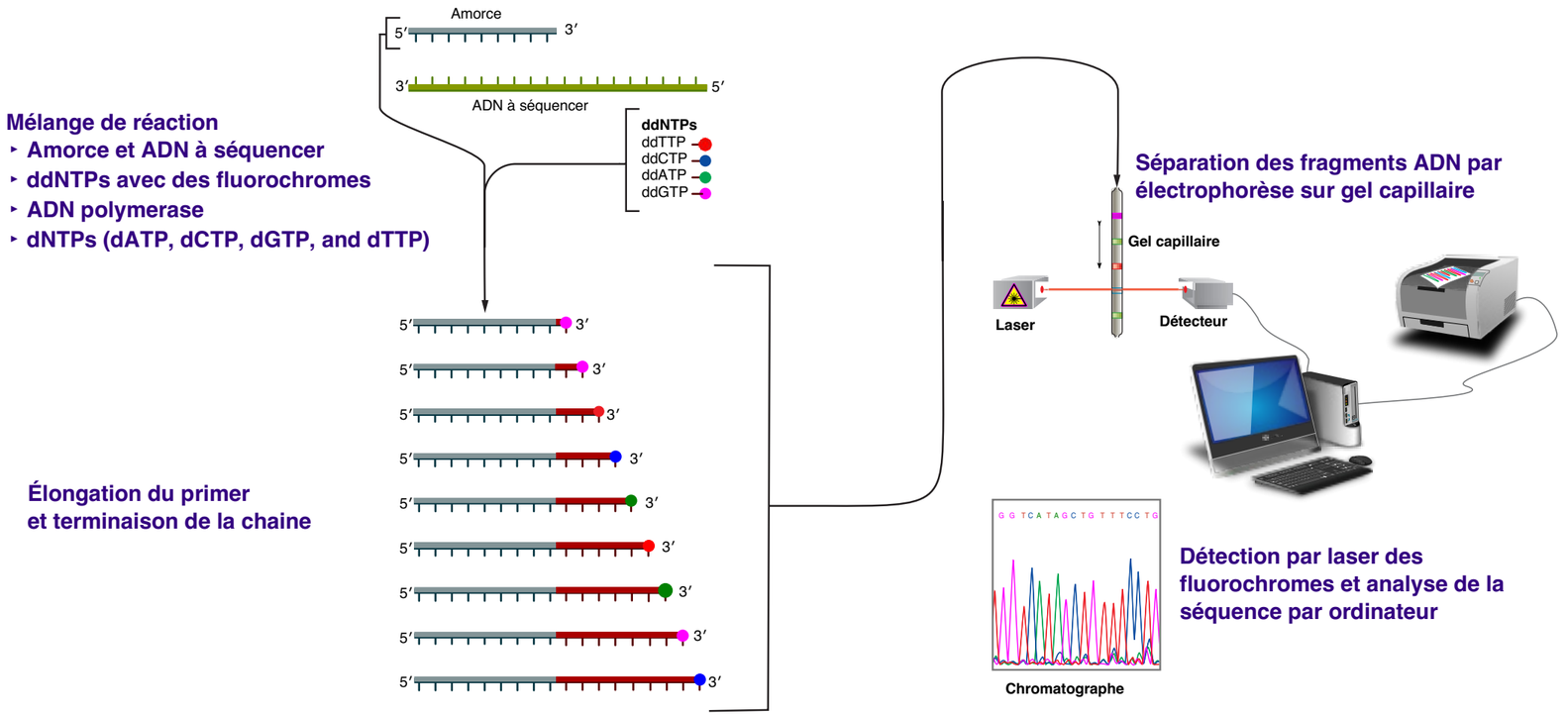
- Revue sur l'histoire du début du séquençage et de la bioinformatique : *Hutchinson, Nucl Acids Res, 2007*
- Revues sur les différentes méthodes de séquençage
  - *Metzker, Genome Res, 2005*
  - *Metzker, Nature Rev Genet, 2010*
  - *Schadt et al, Hum Mol Genet, 2010*

Méthodes de séquençage

Première génération

Séquençage de Sanger

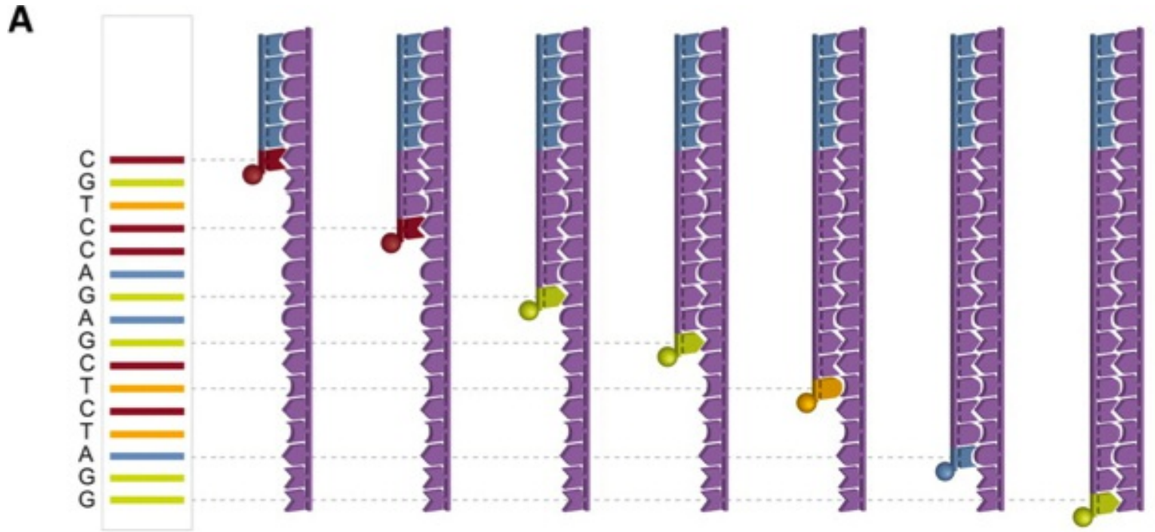
Sanger & Coulson, J Mol Biol, 1975; Sanger et al, Proc Natl Acad Sci USA, 1977



Crédits : Estvezj (modifié)

Séquençage de Sanger

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



Crédits : *Schadt et al, Hum Mol Genet, 2010*

Avantages

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

Limites

- Coûts élevés (~ 3€ par séquence)
- Pas haut-débit (384 séquences par run)

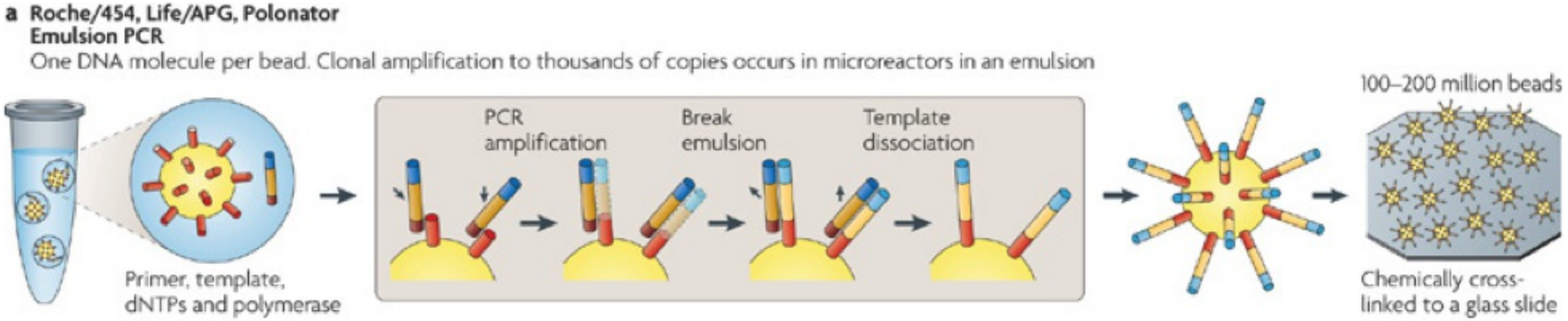
Seconde génération

Principe

- Préparation de l'échantillon
  - Amplification clonale
  - Immobilisation
- Séquençage et visualisation
  - Cycle reversible termination (CRT)
  - Sequencing by ligation (SBL)
  - Single-nucleotide addition (SNA)
  - Real-time sequencing

Préparation de l'échantillon (1)

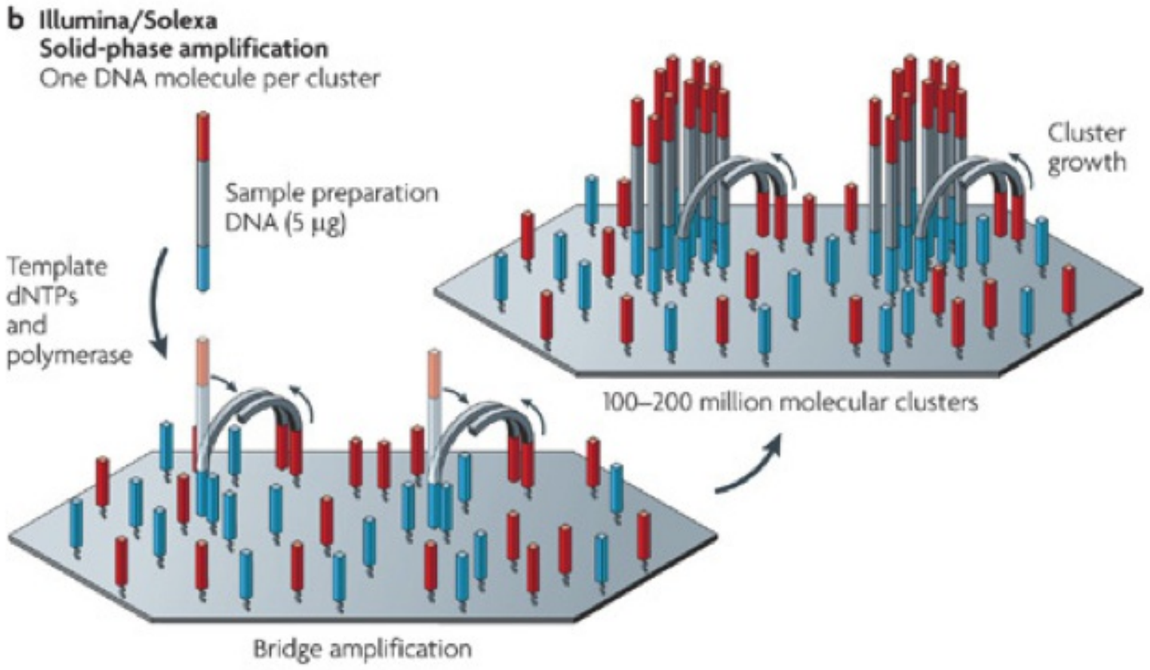
Amplification clonale par émulsion PCR



Crédits : *Metzker, Nature Rev Genet, 2010*

Préparation de l'échantillon (1)

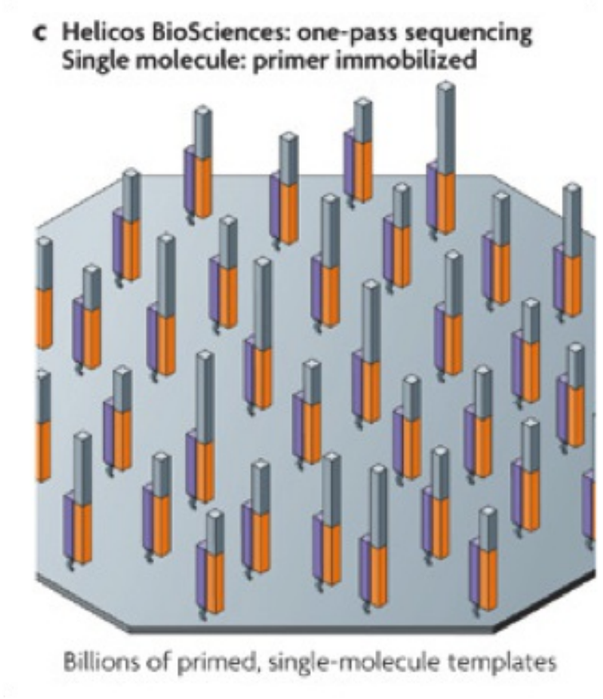
Amplification clonale solid-phase



Crédits : *Metzker, Nature Rev Genet, 2010*

Préparation de l'échantillon (2)

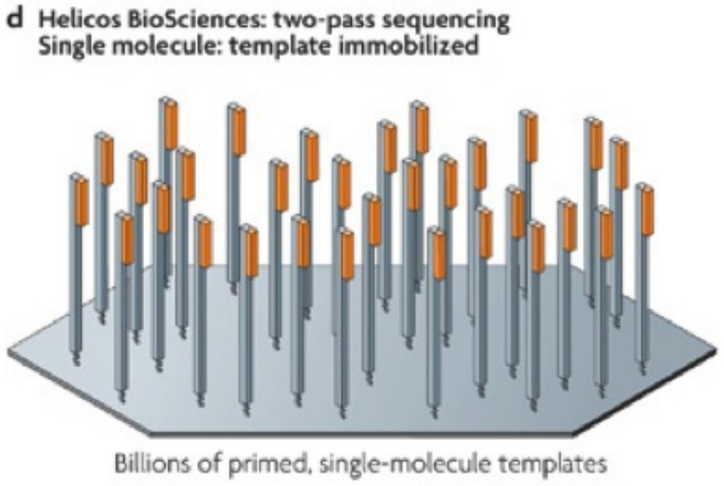
Immobilisation des primers



Crédits : *Metzker, Nature Rev Genet, 2010*

Préparation de l'échantillon (2)

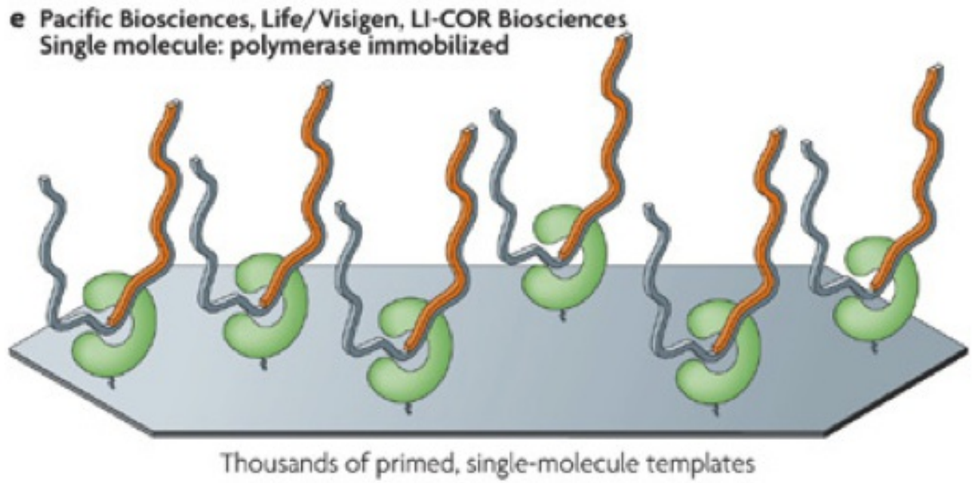
Immobilisation des templates



Crédits : *Metzker, Nature Rev Genet, 2010*

Préparation de l'échantillon (2)

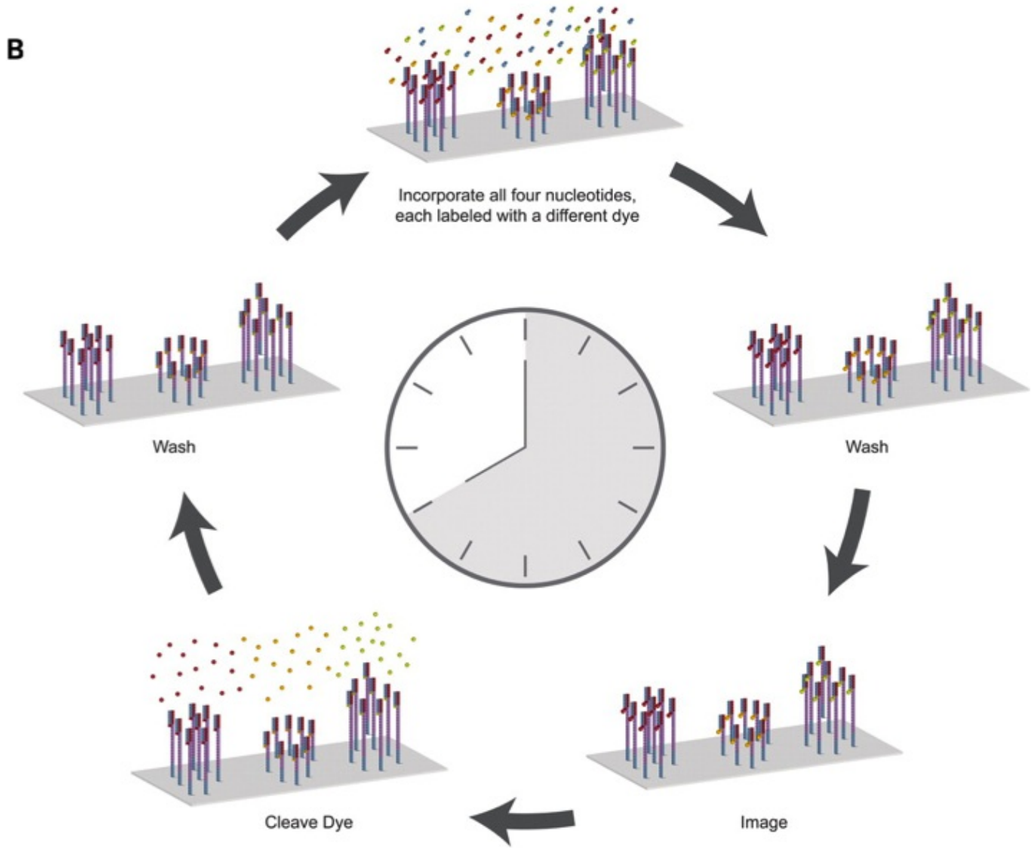
Immobilisation des polymérases



Crédits : *Metzker, Nature Rev Genet, 2010*

Séquençage et visualisation

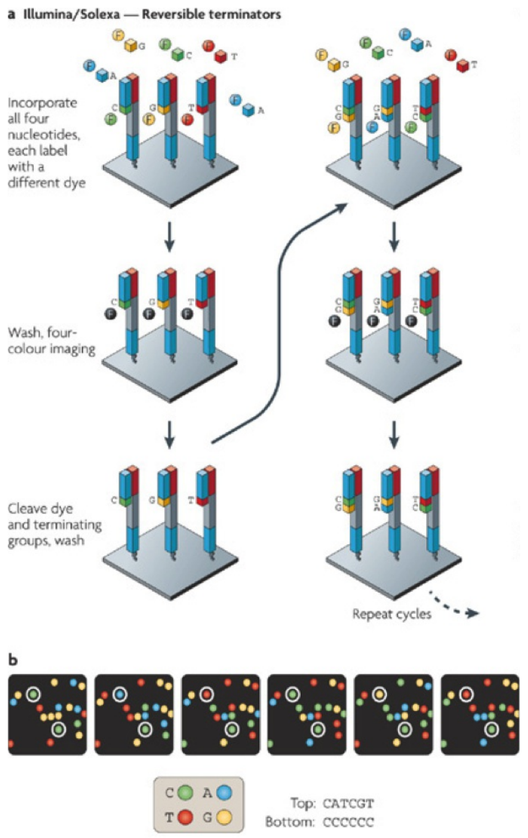
Cyclic reversible termination (CRT)



Crédits : *Schadt et al, Hum Mol Genet, 2010*

Séquençage et visualisation

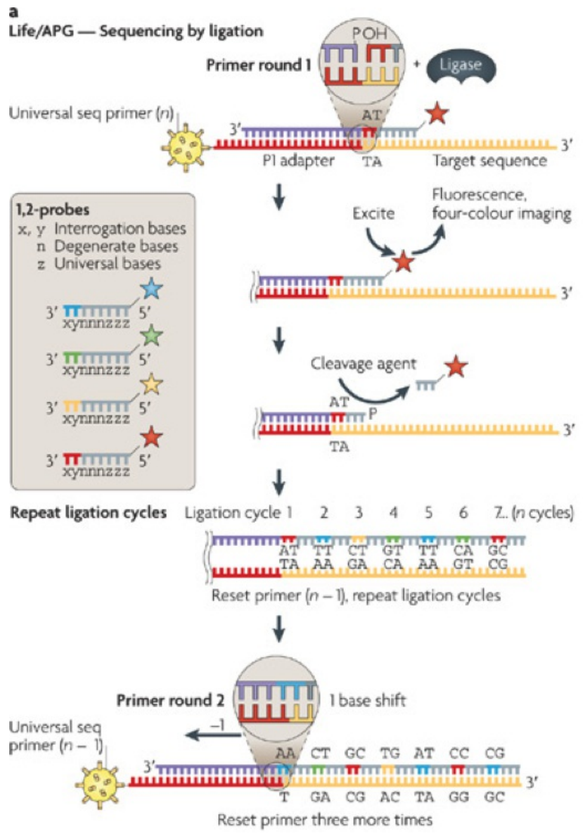
Cyclic reversible termination (CRT)



Crédits : *Metzker, Nature Rev Genet, 2010*

Séquençage et visualisation

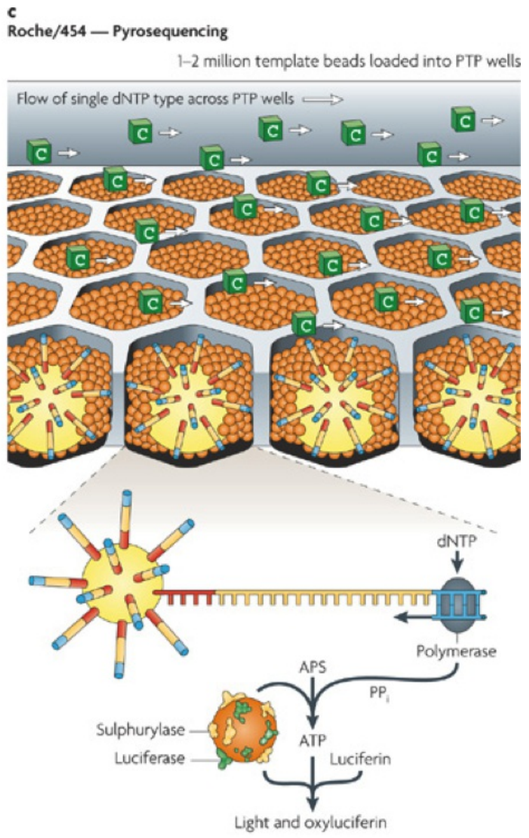
Sequencing by ligation (SBL)



Crédits : *Metzker, Nature Rev Genet, 2010*

Séquençage et visualisation

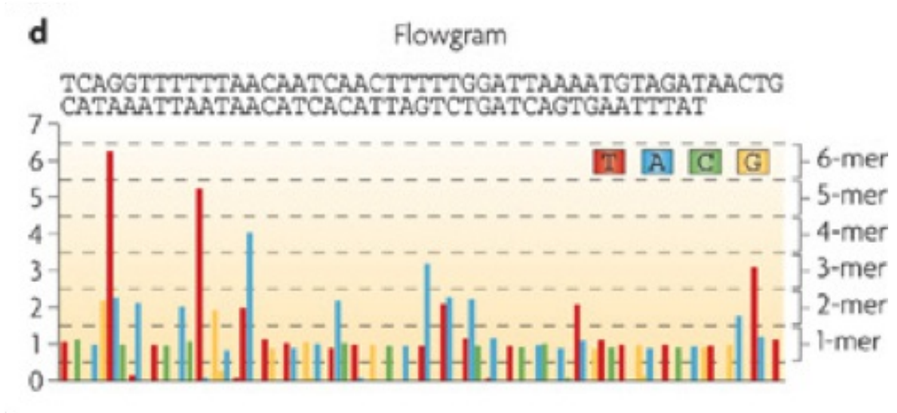
Single-nucleotide addition (SNA) ou pyroséquençage



Crédits : *Metzker, Nature Rev Genet, 2010*

Séquençage et visualisation

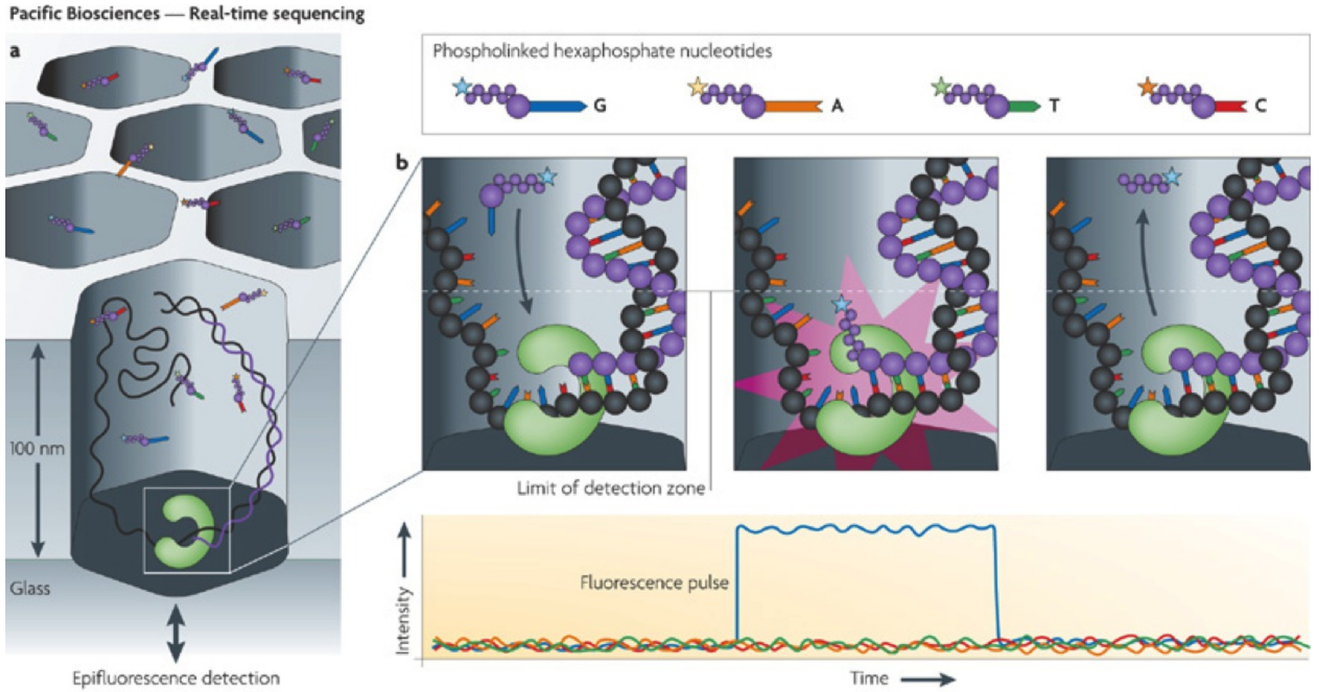
Single-nucleotide addition (SNA) ou pyroséquençage



Crédits : *Metzker, Nature Rev Genet, 2010*

Séquençage et visualisation

Real-time sequencing



Crédits : *Metzker, Nature Rev Genet, 2010*

Différentes plateformes

Plateforme	**Préparation**	**Séquençage**
Roche/454	PCR par émulsion	Pyroséquençage
Illumina	Solid-phase amplification	CRT
Life/APG	PCR par émulsion	SBL
Polonator	PCR par émulsion	SBL
Helicos (Biosciences)	Immobilisation des templates ou primers	CRT
Pacific (Biosciences)	Immobilisation des polymérases	Real-time

Crédits : *Metzker, Nature Rev Genet, 2010*

Illumina

Vidéo YouTube

Avantages

- Haut-débit
- Coûts faibles

Limites

- Amplification
  - Risque d'erreur
  - Augmentation de la complexité et du temps associé à la Préparation
- Génération importante de données
  - Traitement
  - Stockage
- Séquences courtes

Entre la 2e et 3e génération

Ion Torrent

Mesure du largage d'un H+ lors de l'incorporation d'une base avec une puce électronique

Vidéo YouTube

Ion Torrent

- Avantages
  - Pas besoin de lumière, scanning et caméras
- Limites
  - Système "wash-and-scan" avec une amplification PCR

Helicos Genetic Analysis Platform

Imagerie d'ADN individuels fixés sur une surface plane pendant leur élongation avec une polymérase modifiée et des nucléotides fluorescents

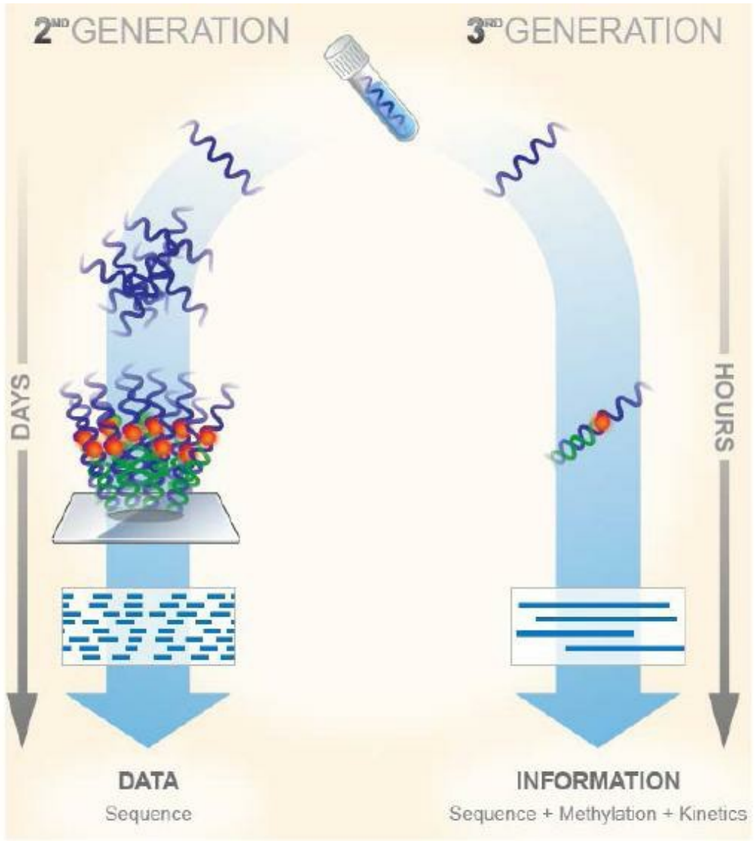
Helicos Genetic Analysis Platform

- Avantages
  - Pas de PCR
- Limites
  - Arrêt de l'élongation
  - Taux d'erreur > 5%

Troisième génération

Troisième génération

Single-molecule sequencing (SMS)



Crédits : http://www.biorigami.com

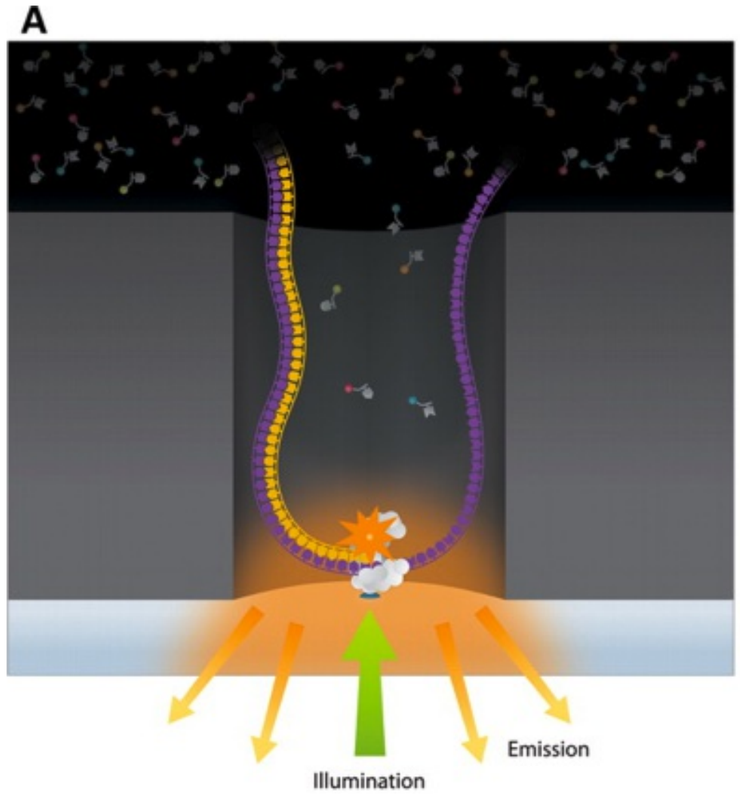
Troisième génération

3 principales méthodes

- SBS
- Nanopores
- Imagerie directe

Sequencing by synthesis (SBS)

Observation de molécules d'ADN polymérases seules lors de la synthèse d'une seule molécule d'ADN



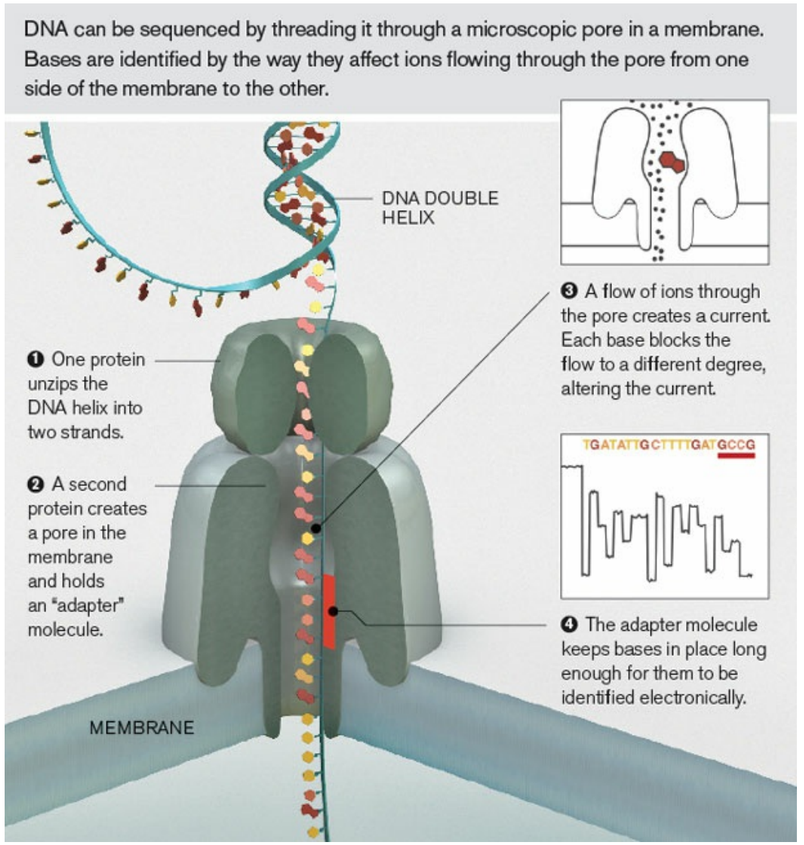
Crédits : *Schadt et al, Hum Mol Genet, 2010*

Sequencing by synthesis (SBS)

2 techniques

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

Nanopores

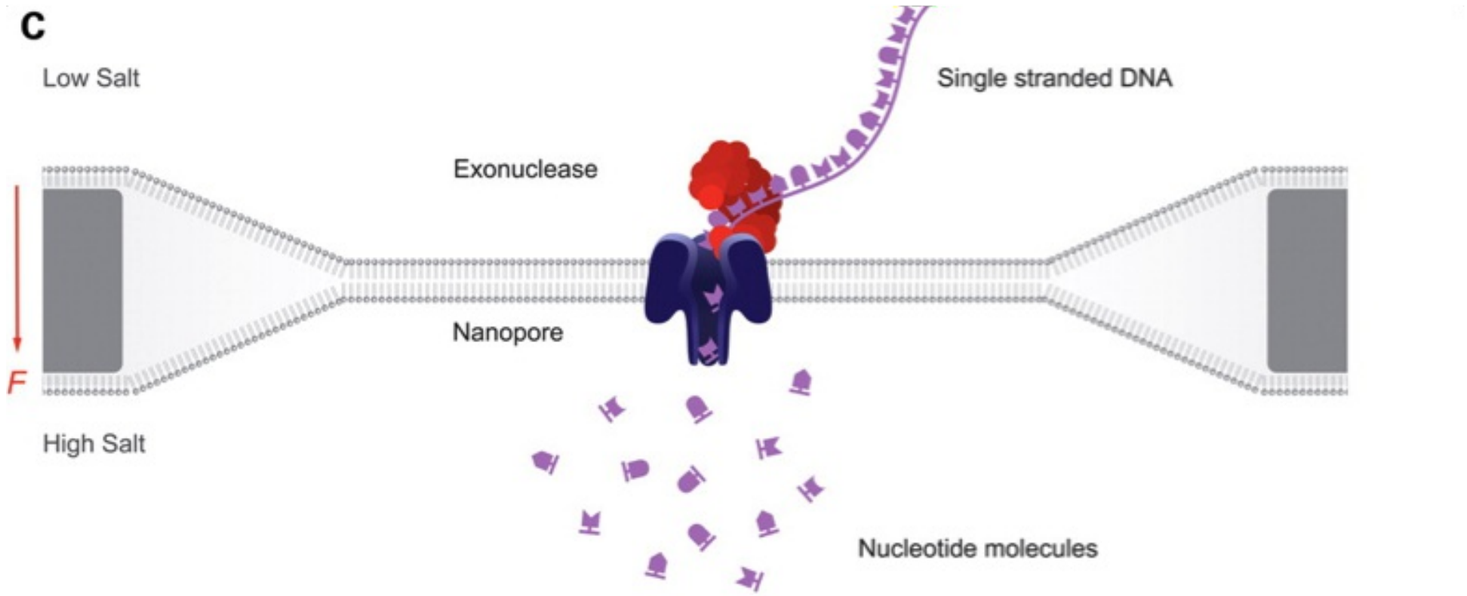


Crédits : John MacNeill

Nanopores

4 méthodes

- Détection électrique directe des molécules d'ADN simple



Crédits : *Schadt et al, Hum Mol Genet, 2010*

Nanopores

4 méthodes

- Détection électrique directe des molécules d'ADN simple

Nanopores

4 méthodes

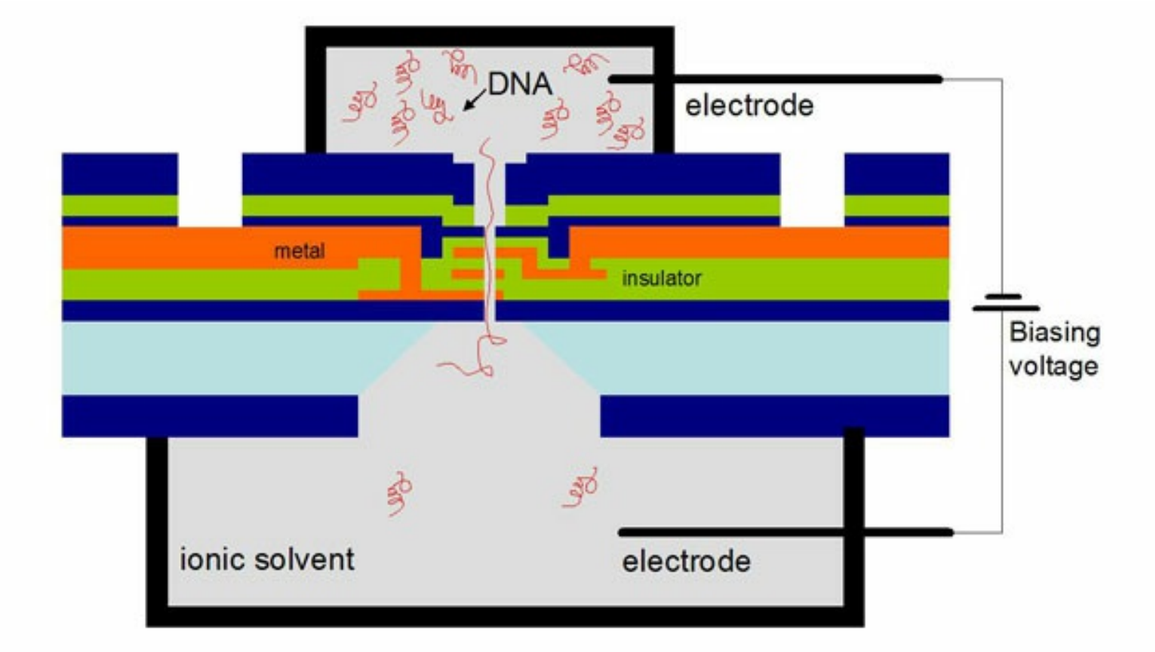
- Détection électrique directe des molécules d'ADN simple
- Séquençage médié par un transistor

Crédits : *Schadt et al, Hum Mol Genet, 2010*

Nanopores

4 méthodes

- Détection électrique directe des molécules d'ADN simple
- Séquençage médié par un transistor

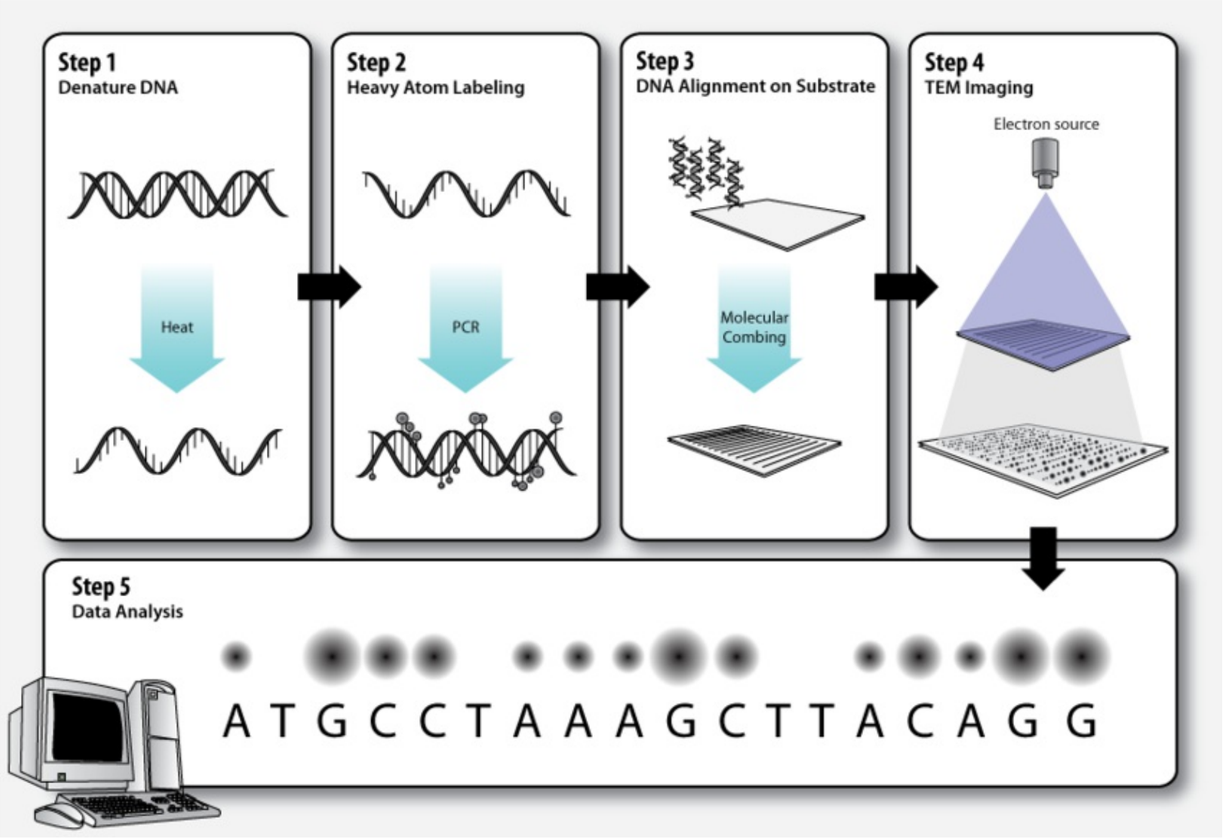


Nanopores

4 méthodes

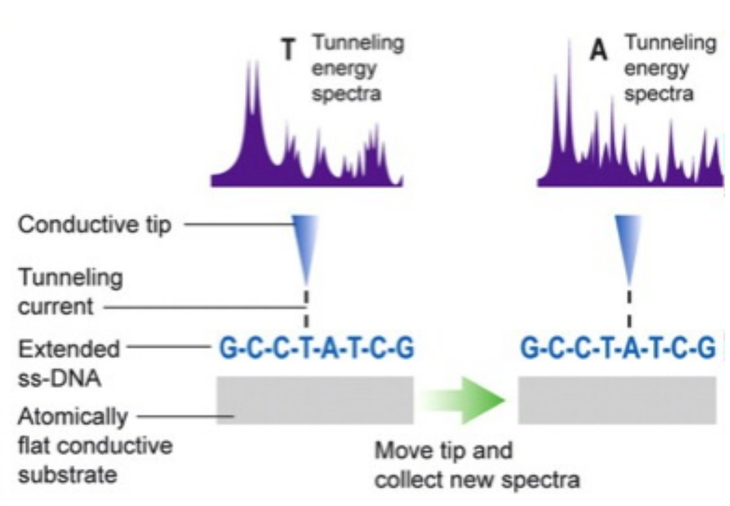
- Détection électrique directe des molécules d'ADN simple
- Séquençage médié par un transistor
- Utilisation d'un nanopore biologique comme MspA (*Mycobacterium smegmatis* Porin A)
- Mesure optique

Imagerie directe des molécules d'ADN individuelles par microscopie



Imagerie directe des molécules d'ADN individuelles par

microscopie



Crédits : *Schadt et al, Hum Mol Genet, 2010*

Troisième génération

Avantages

- Pas besoin de PCR (molécule simple)
- Haut débit
- Délai de production plus courts
- Reads plus longs
- Plus forte précision
- Besoins en ADN plus faibles
- Coûts plus faibles

Comparaison des générations et des méthodes

Comparaison des générations

	1e génération	**2e génération**	**3e génération**
Technologie fondamentale	Discrimination par taille de brin, fragment d'ADN labellisé, produit par SPS ou dégradation	"Wash-and-scan" SPS	SPS, par dégradation, par inspection directe de molécule d'ADN
Résolution	Moyenne selon le nombre de copie de l'ADN séquencé	Moyenne selon le fonction du nombre de copie de l'ADN séquencé	Elevé
Précision	Haute	Haute	Modérée
Taille des séquences	800/ 1000 pb	10/50 pb	1000+ pb

Crédits : *Schadt et al, Hum Mol Genet, 2010*

Comparaison des générations

	1e génération	**2e génération**	**3e génération**
Volume des données	Bas	Haut	Modéré
Coûts	Elevé par base, bas par séquençage	Bas par base, élevé par séquençage	Bas/Modéré par base, bas par séquençage
Temps	~ heures	~ jours	~ heures
Préparation	Modérement complexe, PCR non requise	Complexe, PCR requise	Simple/Complexe suivant la technologie employée
Analyses	Facile	Complexe (quantité de données, taille des reads)	Complexe (quantité de données, types d'information)

Crédits : *Schadt et al, Hum Mol Genet, 2010*

Choix de la plateforme

A 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

Length (bp)	Genome element
25-75	SNP, short frameshift mutation
100-400	Short functional signature
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

Crédits : *Wooley et al, Plos Comp Biol, 2010*

Longueur des reads

- 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 la fréquence (comptage)

Comparaison des plateformes

Instrument	Run time	Millions of Reads/run	Bases / read	Reagent Cost/run	Reagent Cost/Gb	Reagent Cost/Mread	bp/run	Gbp/run	cost/Gb
Applied Biosystems 3730 (capillary)	2 hrs,	0.000096	650	$144	$2307692,31	$1500000,00	624	0	$2307692,31
454 GS Jr, Titanium	10 hrs,	0.1	400	$977	$19540,00	$9770,00	5.0E+07	0.05	$19540,00
454 FLX Titanium	10 hrs,	1	400	$6200	$15500,00	$6200,00	4.0E+08	0.4	$15500,00
454 FLX+	20 hrs,	1	650	$6200	$9538,46	$6200,00	6.5E+08	0.65	$9538,46
Illumina GA IIx - v5 PE	14 days	640	288	$17978	$97,54	$28,09	1.8E+11	184.32	$97,54
Illumina MiSeq v2 Nano	28 hrs,	1	500	$639	$1278,00	$639,00	5.0E+08	0.5	$1278,00
Illumina MiSeq v2 Micro	19 hrs,	4	300	$798	$665,00	$199,50	1.2E+09	1.2	$665,00
Illumina MiSeq v2	39 hrs,	15	500	$1066	$142,13	$71,07	7.5E+09	7.5	$142,13
Illumina MiSeq v3	55 hrs,	22	600	$1442	$109,24	$65,55	1.3E+10	13.2	$109,24
Illumina NextSeq 500	30 hrs,	400	300	$4000	$33,33	$10,00	1.2E+11	120	$33,33
Illumina HiSeq 2500 - rapid run	40 hrs,	300	300	$4126	$45,84	$13,75	9.0E+10	90	$45,84
Illumina HiSeq 2500 - high output v3	11 days	1500	200	$13580	$45,27	$9,05	3.0E+11	300	$45,27
Illumina HiSeq 2500 - high output v4	6 days	2000	250	$14950	$29,90	$7,48	5.0E+11	500	$29,90
Illumina HiSeq X (2 flow cells)	3 days	6000	300	$12750	$7,08	$2,13	1.8E+12	1800	$7,08
Ion Torrent – PGM 314 chip	3,7 hrs,	0.475	400	$474	$2494,74	$997,89	1.9E+08	0.19	$2494,74
Ion Torrent – PGM 316 chip	4,9 hrs,	2.5	400	$674	$674,00	$269,60	1.0E+09	1	$674,00
Ion Torrent – PGM 318 chip	7,3 hrs,	4.75	400	$874	$460,00	$184,00	1.9E+09	1.9	$460,00
Ion Torrent - Proton I	4 hrs,	70	175	$1000	$81,63	$14,29	1.2E+10	12.25	$81,63
Ion Torrent - Proton II (forecast)	5 hrs,	280	175	$1000	$20,41	$3,57	4.9E+10	49	$20,41
Ion Torrent - Proton III (forecast)	6 hrs,	500	175	$1000	$11,43	$2,00	8.8E+10	87.5	$11,43
Life Technologies SOLiD – 5500xl	8 days	1410	110	$10503	$67,72	$7,45	1.6E+11	155.1	$67,72
Pacific Biosciences RS II	2 hrs,	0.03	3000	$100	$1111,11	$3333,33	9.0E+07	0.09	$1111,11
Oxford Nanopore MinION (forecast)	≤6 hrs,	0.1	9000	$900	$1000,00	$9000,00	9.0E+08	0.9	$1000,00
Oxford Nanopore GridION 2000 (forecast)	varies	4	10000	$1500	$37,50	$375,00	4.0E+10	40	$37,50
Oxford Nanopore GridION 8000 (forecast)	varies	10	10000	$1000	$10,00	$100,00	1.0E+11	100	$10,00

Crédits : The Molecular Ecologist

Multiplexage

Moyen de diminuer les coûts de séquençage

- Ajout d'un tag d'identification pour chaque échantilon
- Mélange des échantillons
- Préparation et séquençage en parallèle
- Tri des séquences en fonction des échantillons (information de source contenue dans la séquence)

Taux d’erreur

Instrument	Erreurs principales	Taux d'erreur (%s)
454	Indels	1
Illumina	Substitutions	0.1
Ion Torrent	Indels	1
SOLiD	Biais AT	<= 0.1
Oxford Nanopore	Délétions	4
PacBio RS	Indels	<= 1

Crédits : The Molecular Ecologist

Comparaison des plateformes

Instrument	Avantages	Inconvénient
454 FLX+	Reads contigus les plus longs des 2nd génération	Coûts éleveés par Mb, problèmes de reagents
454 GS Jr	Plus faibles coûts par run que 454 FLX+	Coûts élevés par Mb, Peu de reads, Reads plus courts que FLX
Illumina HiSeq	Failes coûts par Mb	Coûts élevés d'instruments et de runs; Besoin de personnels qualifiés
Illumina MiSeq	Coûts modérés d'instruments et runs, Faibles coûts par Mb pour une petite plateforme, Durée des runs les rapides et les reads les plus longs des plateformes Illumina	Relativement peu de reads et coûts plus élevés par Mb que HiSeq
Illumina NextSeq	Facilité d'utilisation, coûts modérés d'instruments et de run	Nouvelle chimie et nouveau design d'instruments

Crédits : The Molecular Ecologist

Comparaison des plateformes

Instrument	Avantages	Inconvénient
Ion Torrent	Faibles coûts d'instruments, machine simple	Plus forts taux d'erreurs qu'Illumina, plus de temps de manipulation, moins de reads à coûts plus élevés que MiSeq
PacBio	Séquençage en temps réel simple molécule, Reads plus longs; Capacité de détection des modifications de bases	Forts taux d'erreur, Faible nombre de reads par run, Coûts élevés par Mb
SOLiD	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	Instrument portable et low-cost, Reads extrêmement longs	Très récent

Crédits : The Molecular Ecologist

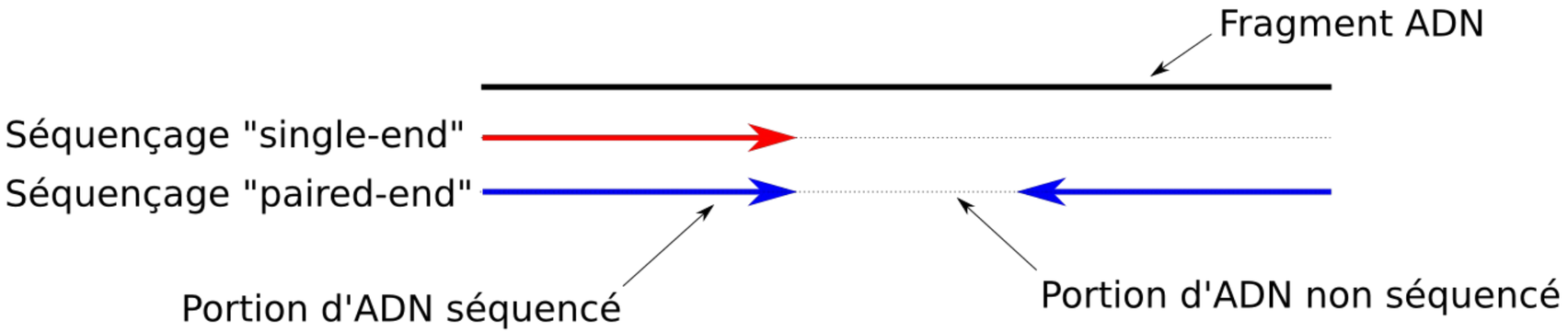
Traitement des données issues du séquençage

Traitement des données issues du séquençage

- Gestion des données
- Prétraitement des données
- Contrôle de la qualité des séquences
- Assemblage
- Analyses des séquences

Gestion des données

Types de données



Types de données

- Séquençage single-end
  - 1 fichier de sortie
- Séquençage paired-end
  - 2 fichiers de sortie
  - Concaténation des fichiers en faisant attention au sens de lecture

Formats des données brutes

- FastA: Format répandu de stockage des séquences biologiques
- FastQ (*Illumina*)
- SFF (*Roche 454*)
- SRF (*Helicos*)
- HDF5 (*PacBio*, *Applied Biosystems*, *Oxford Nanopore*)

FastA

```
> Identifiant (Commentaire)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
> Identifiant2 (Commentaire)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XX
```

FastA

Pas de standard dans les description malgré des tentatives

Banque de données	Description
GenBank	`gi|numero_gi|gb|numero_accession|locus`
EMBL	`gi|numero_gi|emb|numero_accession|locus`
DDDJ	`gi|numero_gi|dbj|numero_accession|locus`
NBRF PIR	`pir||entree`
Protein Research Foundation	`prf|nom`
Swiss-Prot	`sp|numero_accession|nom`
Brookhave PDB	`pdb|entree|chaine`
Brevets	`pat|brevet|numero`
GenInfo Backbone Id	`bbs|numero`
General datase identifier	`gnl|base_de_donnees|identifiant`
NCBI Reference Sequence	`ref|numero_accession|locus`
Local Sequence idenfier	`lci|identifiant`

Code des acides nucléoiques

Code	Signification	Mémorisation
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	*K*etone (cétone)
M	A ou C	a**M**ine
S	C ou G	**S**trong (interaction forte)
W	A, T ou U	**W**eak (interaction faible)

Code des acides nucléoiques

Code	Signification	Mémorisation
B	différent de A	**B** vient après A
D	différent de C	**D** vient après C
H	différent de G	**H** vient après G
V	différent de T ou U	**V** vient après U
N	A, C, G, T ou U	**N**'importe
X	acide nucléique masqué	
-	gap	

Code des acides aminés

Standard IUB/IUPAC

Code	Signification	Code	Signification
A	Alanine	O	Pyrrolisine
B	Acide aspartique	P	Proline
C	Cystéine	Q	Glutamine
D	Acide aspartique	R	Arginine
E	Acide glutamique	S	Sérine
F	Phénylalanine	T	Thréonine
G	Glycine	U	Sélénocystéine
H	Histidine	V	Valine
I	Isoleucine	W	Tryptophane
K	Lysine	Y	Tyrosine
L	Leucine	Z	Acide glutamique
M	Méthionine	X	N'importe
N	Asparagine	*	Codon stop

Différents fichiers en FastA

Extension	Signification	Commentaire
`.fasta`, `.fas`, `.fa`	FastA générique	Tout fichier fasta
`.fna`	FastA nucleic acid	Fichier fasta contenant une séquence d'acides nucléiques
`.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
`.frn`	FastA non-coding RNA	Fichier fasta contenant une séquence d'ARN non-codant d'un génome

FastQ

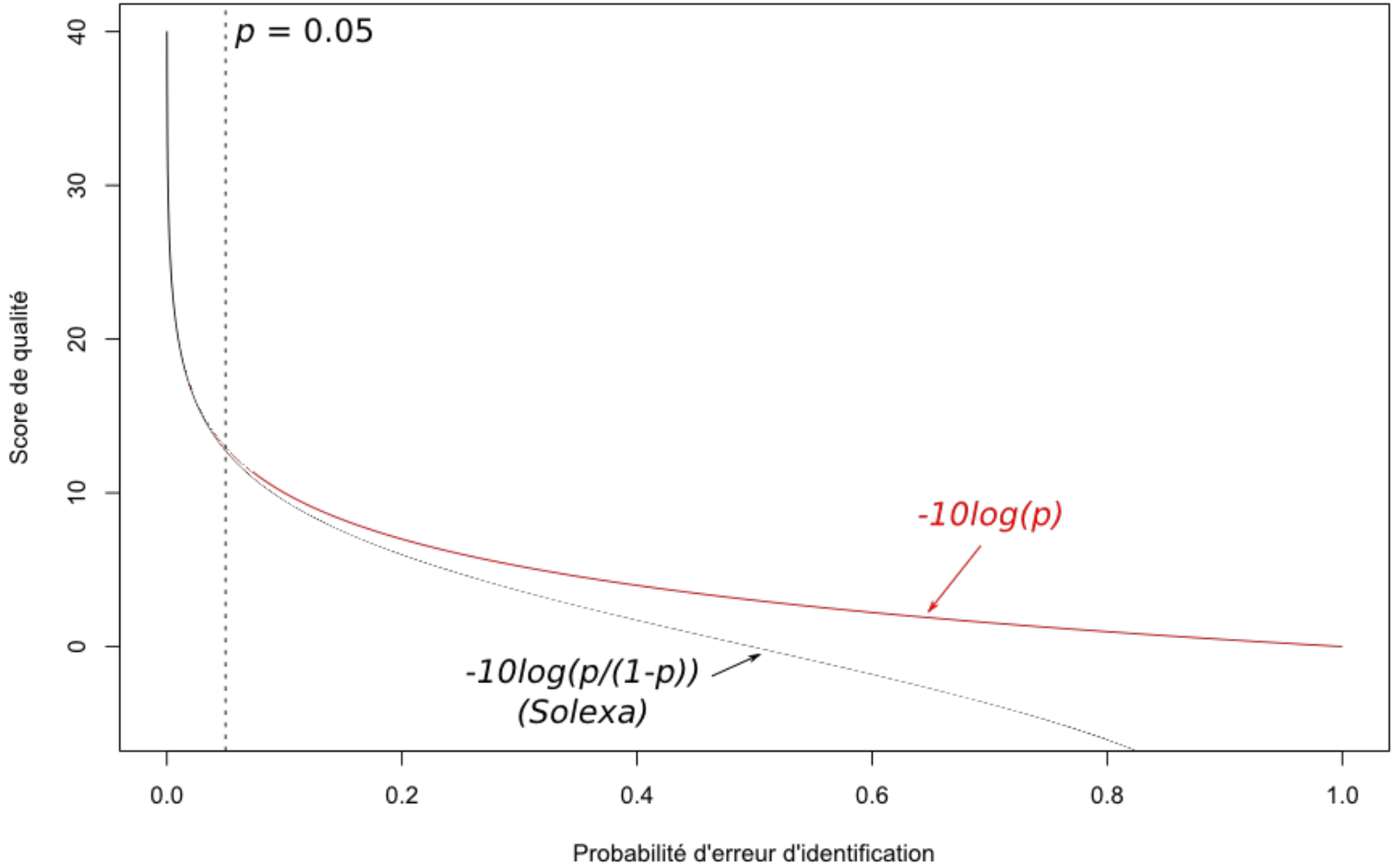
```
> Identifiant (Commentaire)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
+
QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
> Identifiant2 (Commentaire)
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
+
QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
```

Score de qualité

Mesure de qualité de l'identification des bases par le séquençage ADN

Score assigné à chaque base dans les sorties des séquenceurs

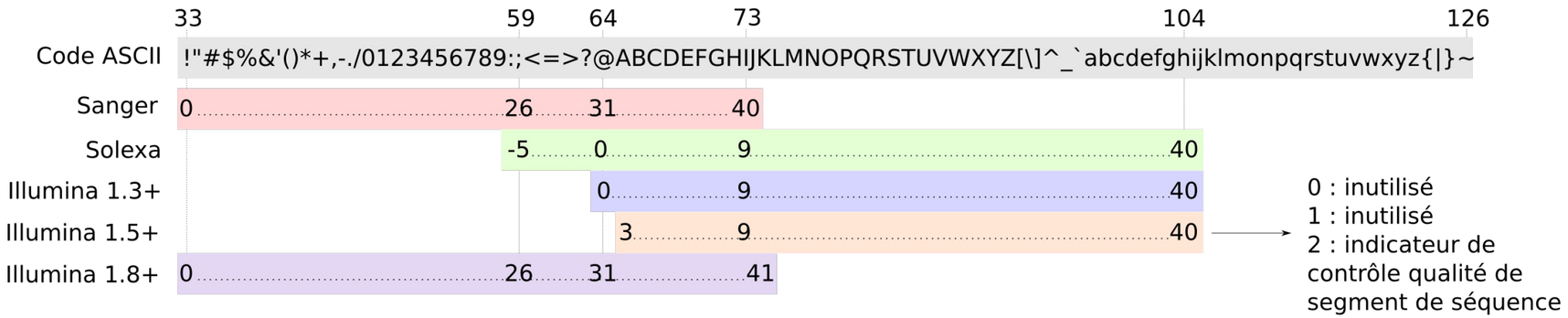
Score de qualité



Score de qualité

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 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

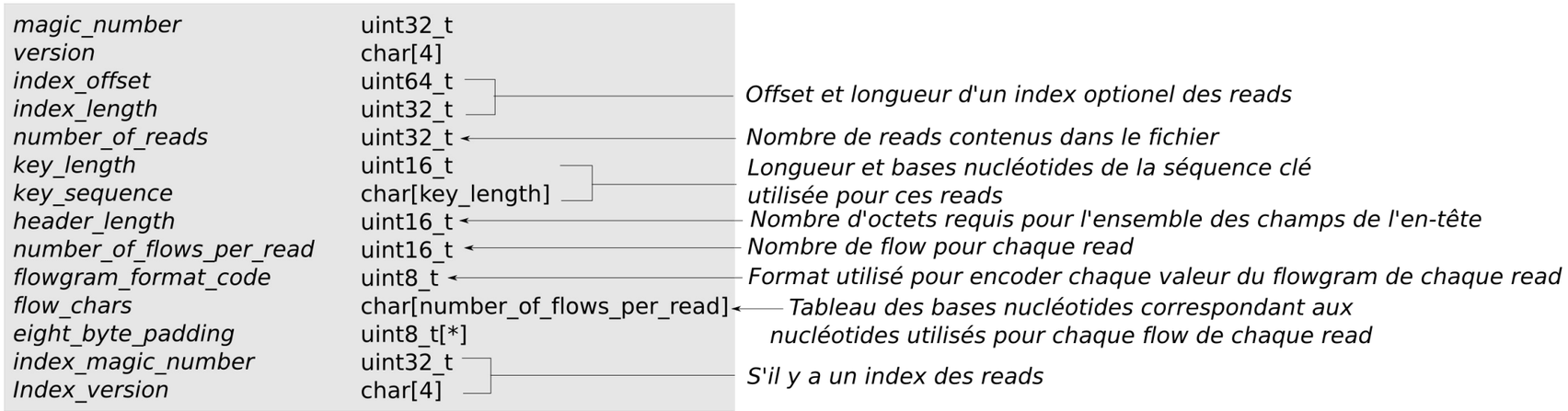
Codage du score de qualité



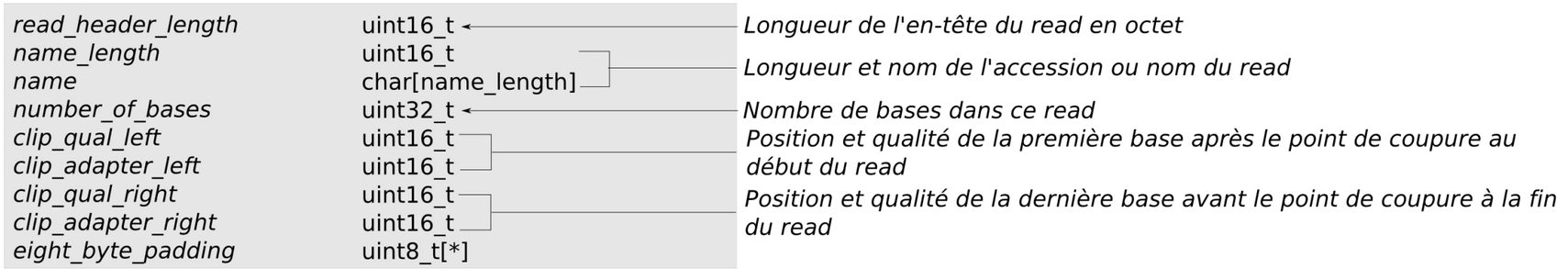
Standard Flowgram Format (SFF)

- Format de sortie de 454 et Ion Torrent
- Fichier binaire
  - Section en-tête commune à tous les reads
  - Pour chaque read
    - Section en-tête du read
    - Section des données du read

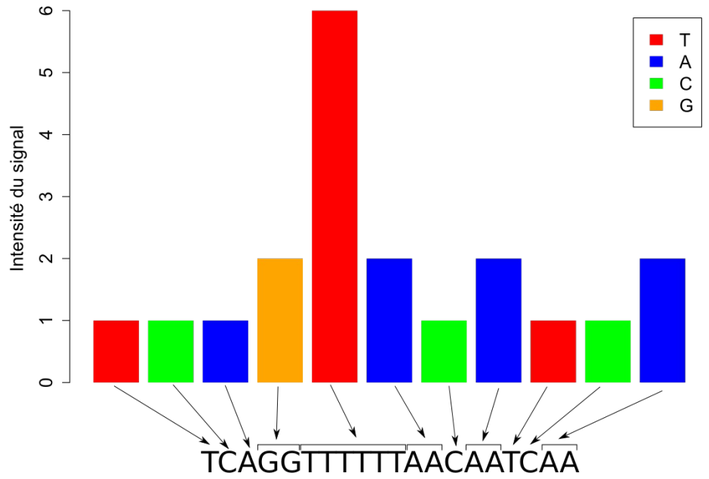
En-tête globale d'un fichier SFF



En-tête pour un read dans un fichier SFF



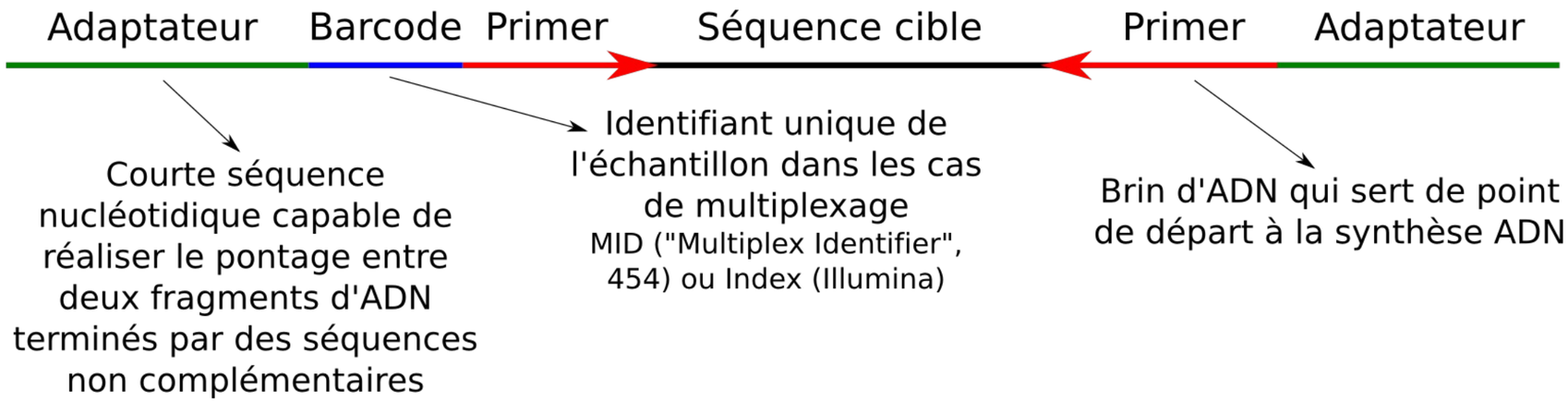
Données d'un read dans un fichier SFF



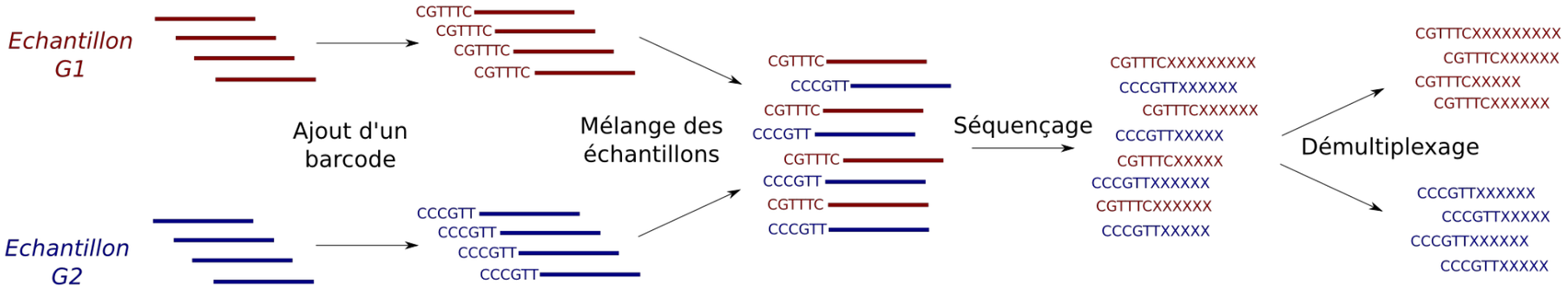
flowgram_values	uint*_t[number_of_flows]	*Estimation de l'étirement de l'homopolymère pour chaque flow*
flow_index_per_base	uint8_t[number_of_bases]	*Position de flow de chaque base dans la séquence*
bases	char[number_of_bases]	*Séquence de nucléotides*
quality_scores	uint8_t[number_of_bases]	*Score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

Prétraitement des données

Primers, adaptateurs, barcodes



Multiplexage / Démultiplexage



Démultiplexage

Utilisation de fichier de "mapping"

Exemple :

Sample ID	Barcode	Primer	Description
G1	CGTTTC	ACGGRAGGCAGGCAG	Crohn DNA
G2	CCCGTT	ACGGRAGGCAGGCAG	Healthy DNA
G3	GGTCAC	ACGGRAGGCAGGCAG	Crohn RNA
G4	AGTGCT	ACGGRAGGCAGGCAG	Healthy RNA

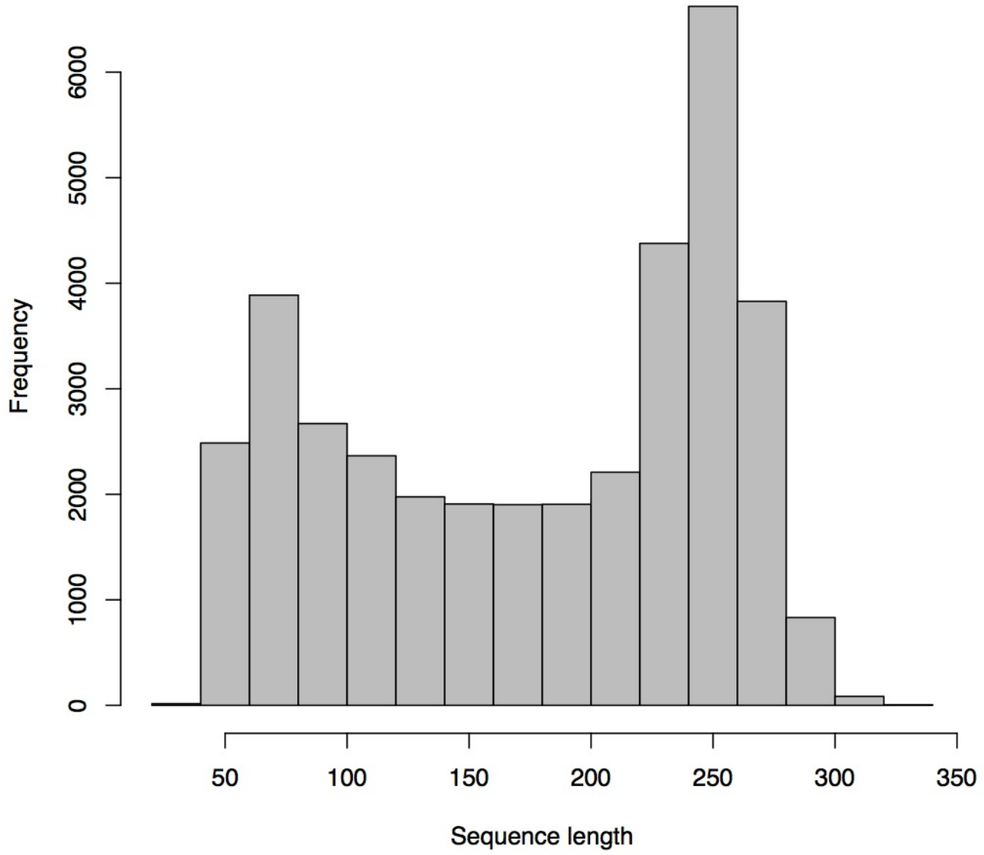
Contrôle de la qualité des séquences

Contrôle de la qualité des séquences

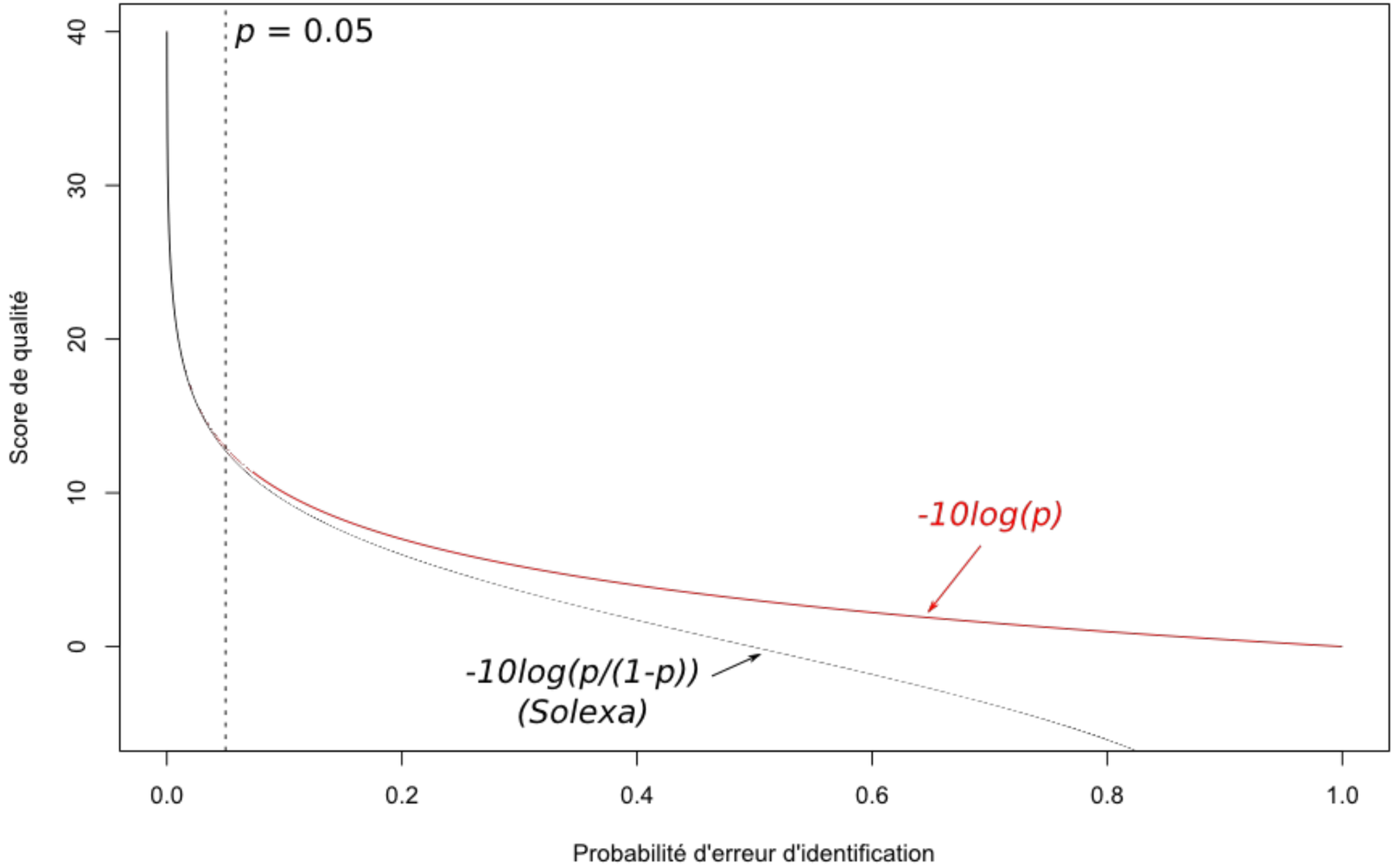
Plusieurs paramètres à vérifier

- Longueur des séquences
- Score de qualité des bases
- Contenu en bases
- Duplications de séquence
- Complexité des séquences
- Contamination

Longueur des séquences



Score de qualité des bases

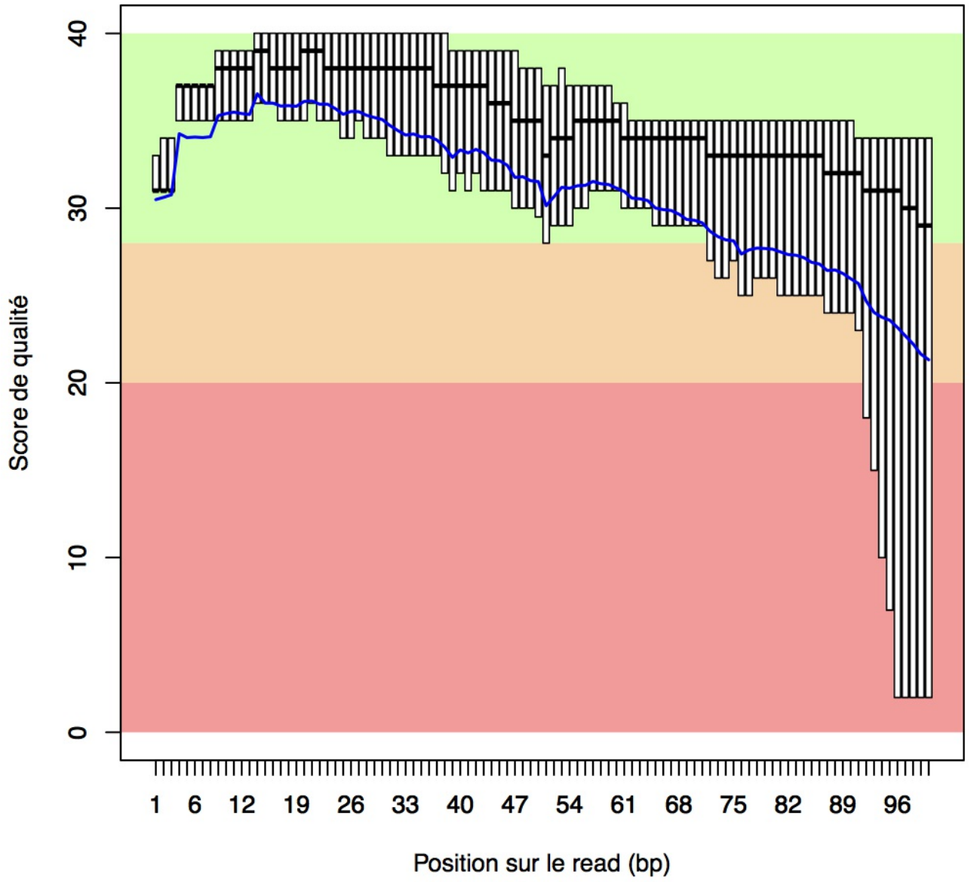


Score de qualité des bases

2 niveaux à vérifier

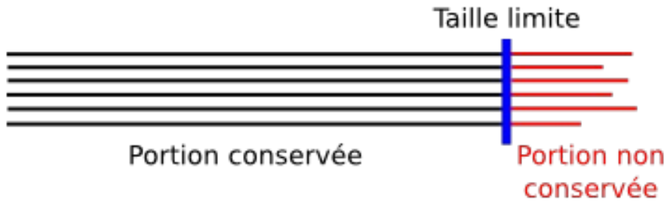
- Score de qualité de la séquence entière
  - Eliminer les séquences avec score moyen < 25
  - Huse et al, Genome Biology, 2007
- Score de qualité par base

Score de qualité par base



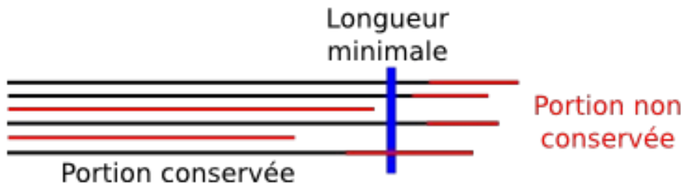
Coupure des séquences basées sur le score de qualité par base

Coupure de longueur fixe

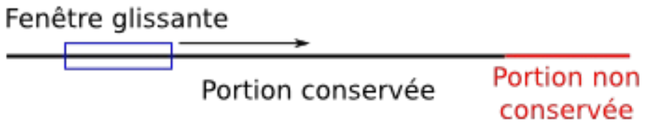


Coupure des séquences basées sur le score de qualité par base

Coupure de longueur variable



Utilisation d'une fenêtre glissante

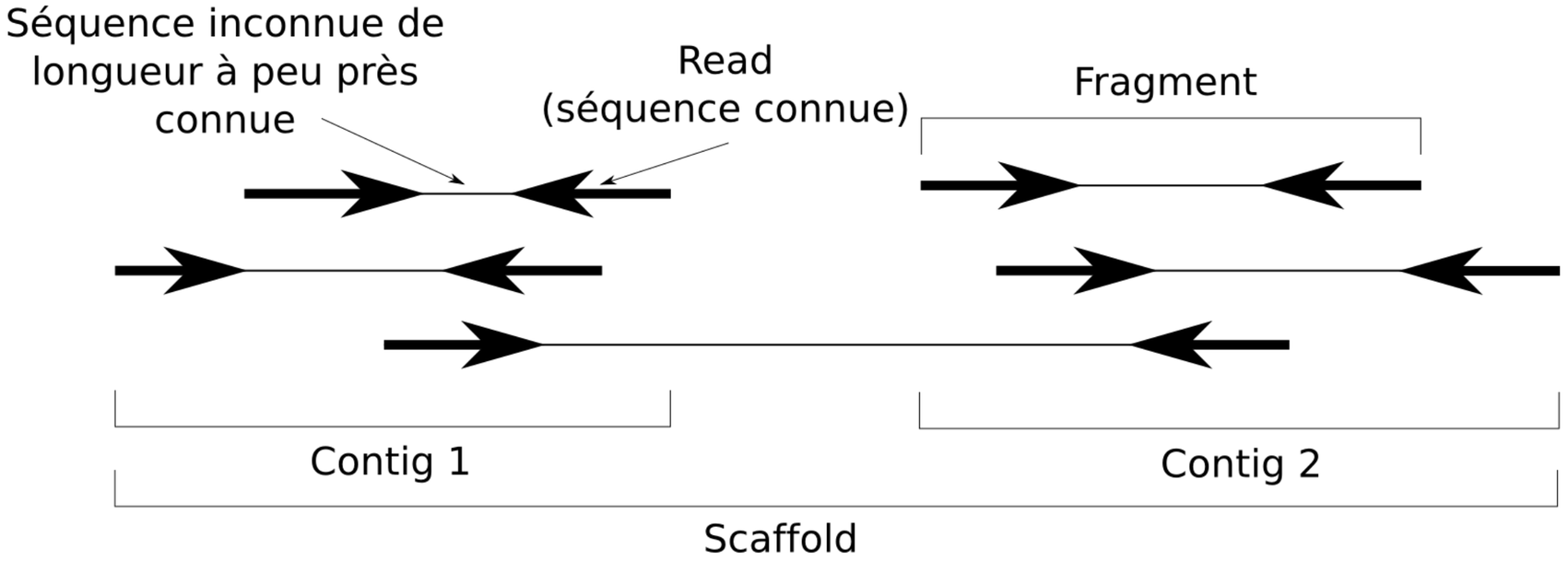


Assemblage

Terminologie

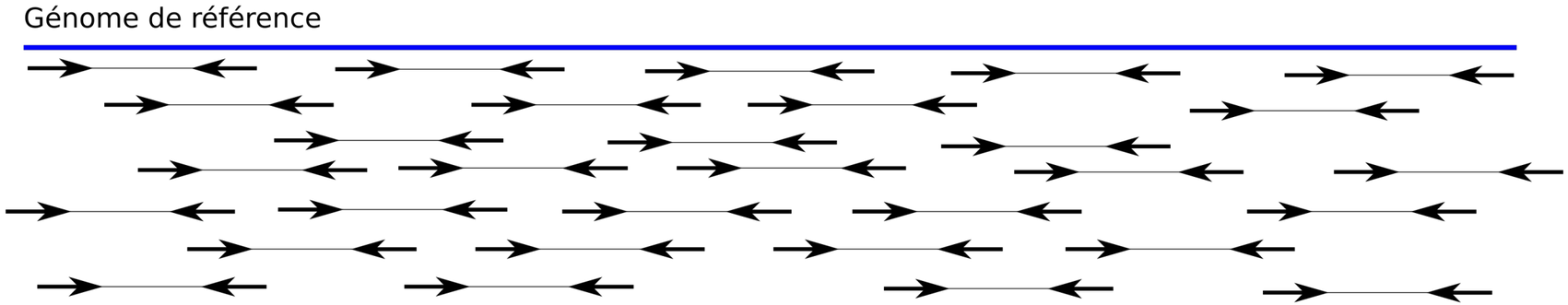
- **Assemblage** : Alignement et fusion de reads en séquences ADN plus longues
- **Contig** : Séquence génomique continue et ordonnée générée par l'assemblage de reads qui se chevauchent
- **Scaffold** : Contigs chevauchant séparés par des gaps de longueur connue

Terminologie



Utilisation de l'assemblage

Alignement sur un génome de référence



- Identification de variants
- Reséquençage de génomes

Utilisation de l’assemblage

Assemblage *de novo*

Construction des séquences ADN d'un organisme sans 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 aléatoire
- **Profondeur de couverture** : Nombre de reads x Longueur des reads / Taille de l'assemblage

Métriques d'assemblage

- Nombre de contigs/scaffolds
- Taille des contigs/scaffolds
- Taille totale
- Nombre de N
- N50
  - Longueur du plus petit contig dans l'ensemble qui contient le moins de configs dont les longueurs combinées représente au moins 50% de l'assemblage

Analyse des séquences

Analyse des séquences

- Construction de génomes
- Détection des SNP ou indels
- Détection de nouveaux gènes ou éléments régulateurs
- Détermination des niveaux d'expression
- ...

Références

- Quelques vidéos sur les différentes méthodes de séquençage
  - Apport d'Ion Torrent par rapport à Illumina

Métagénomique