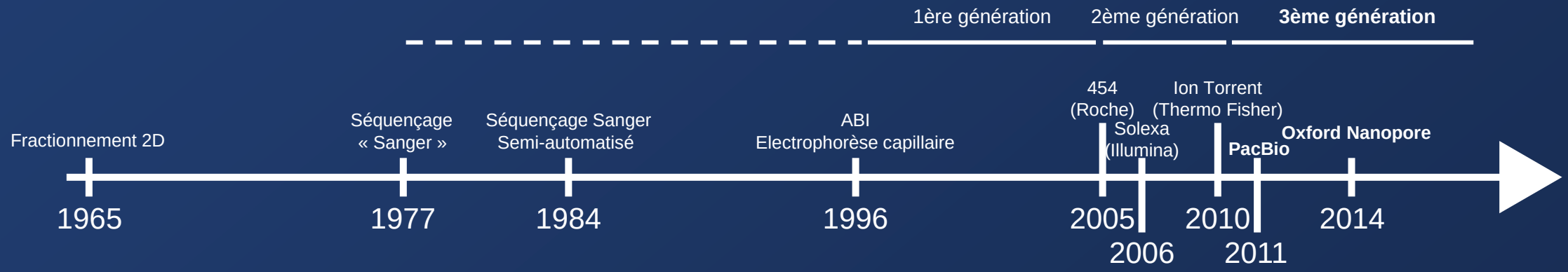
GenoScreen
Innovative Genomics

Les apports du séquençage long read en génomique

Téo Fournier
09/03/2022

La maîtrise de l'information génomique au service de l'Homme et de son environnement

1. Les technologies « long read »



GenoScreen
Innovative Genomics

Comparaison des différentes générations de séquençage

	« Sanger »	Illumina / Thermo Fisher	Pacific Biosciences / Oxford Nanopore Technologies
			
	1ère génération	**2ème génération**	**3ème génération**
Librairie hétérogène (shotgun sequencing)	Non	Oui	Oui
Séquençage direct (Single molecule)	Non	Non	Oui
Qualité des reads	Haute	Haute	**Moyenne** Haute (PacBio Hifi)
Longueur max des reads	**900 pb**	**300 pb**	**100 kb** (PacBio) **>1Mb** (ONT)

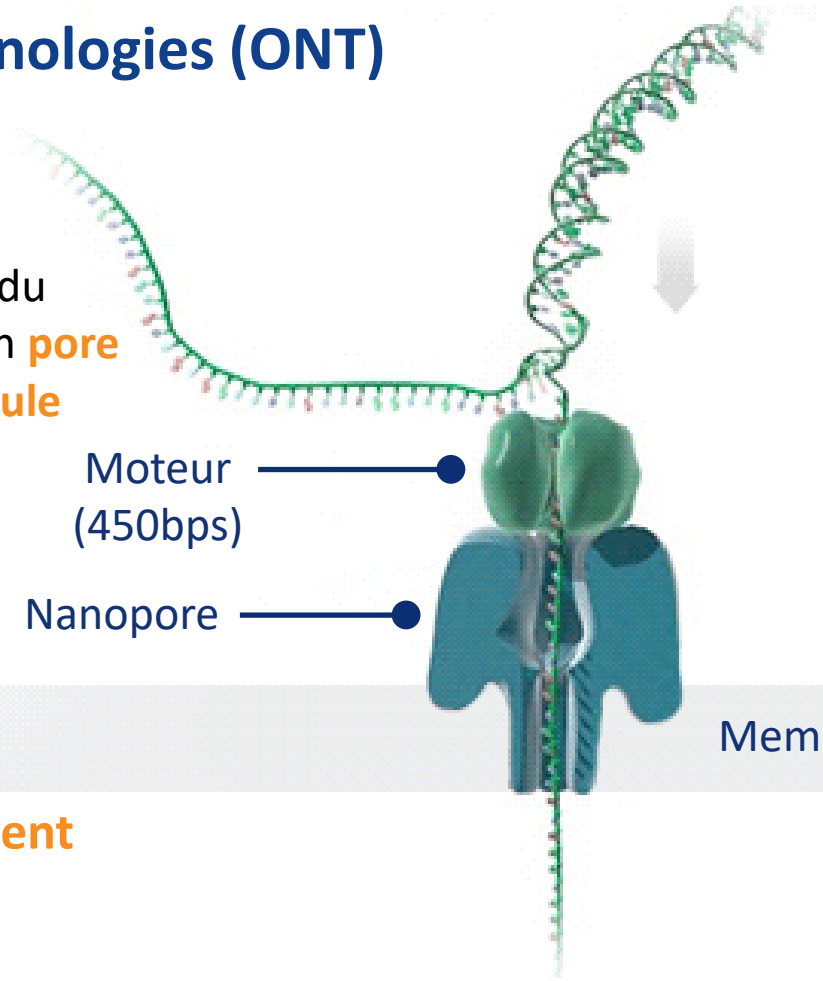
Le séquençage de 3ème génération

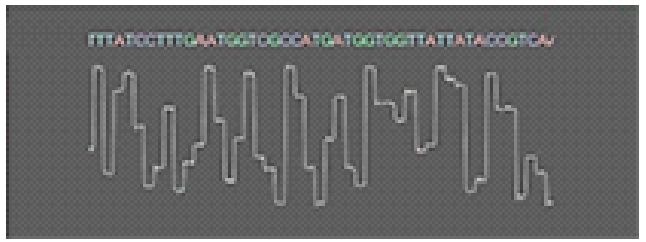
Oxford Nanopore Technologies (ONT)

Principe :
Mesure du changement d'intensité du **courant ionique** passant à travers un **pore** lors de la translocation **d'une molécule** d'ADN (ou ARN)

Qualité des lectures :
Médiane à 95%
Mais… nouvelle chimie à **99%**

Longueur des reads **théoriquement** illimitée





Séquençage en direct

Source : Oxford Nanopore Technologies

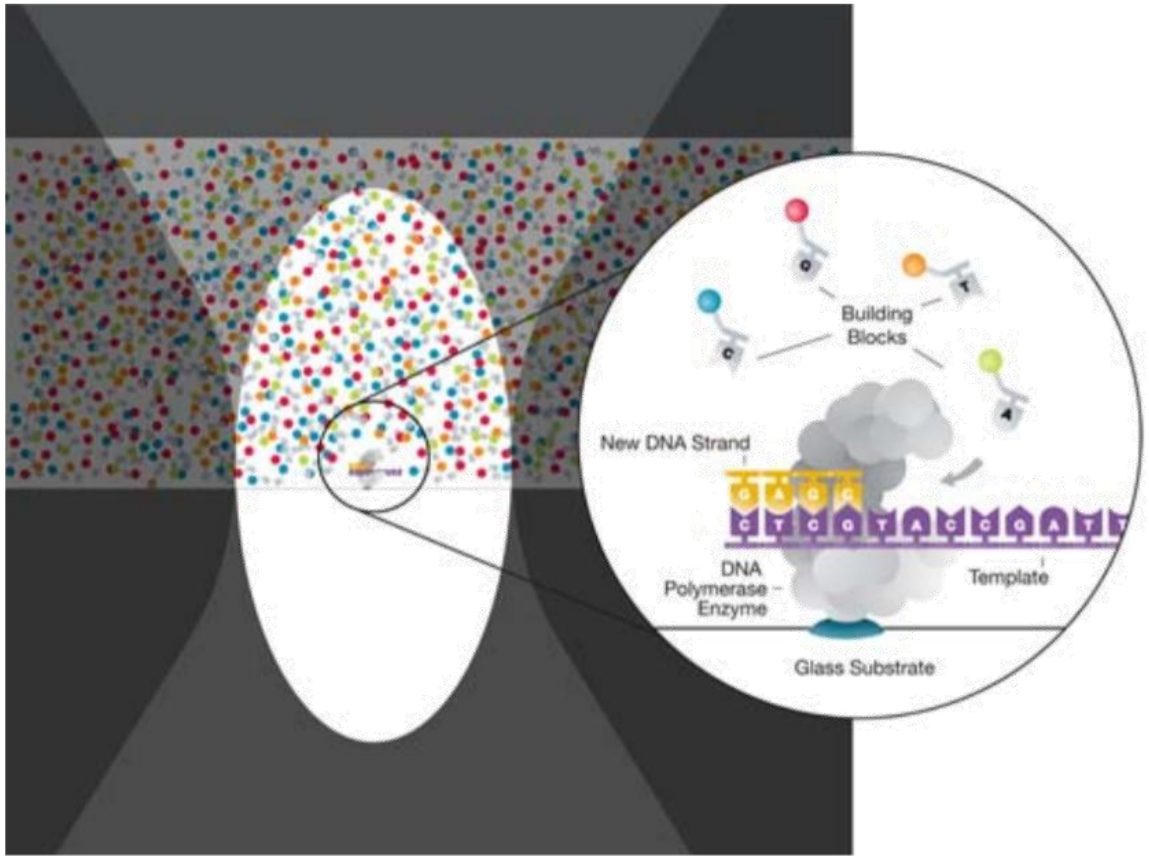
Le séquençage de 3ème génération

Pacific Biosciences (PacBio)

Principe :

ADN polymérase fixée au fond d'un puit

Détection de la fluorescence à chaque insertion de nucléotide



Source : https://www.ndsu.edu/pubweb/~mcclean/plsc411/Pacific%20Biosciences-technology_backgrounder.pdf

Le séquençage de 3ème génération

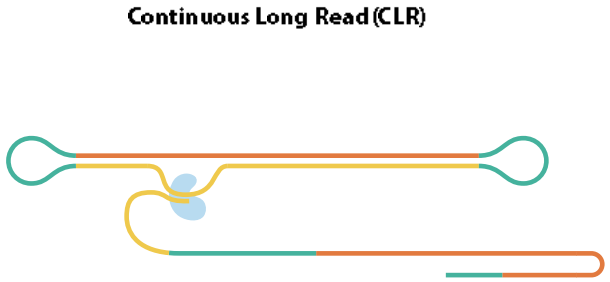
Pacific Biosciences (PacBio)

Deux modes de fonctionnement :

Continuous **L**ong **R**ead (CLR) :

Obtenir les plus **longs** fragments possible, **une seule lecture** par fragment

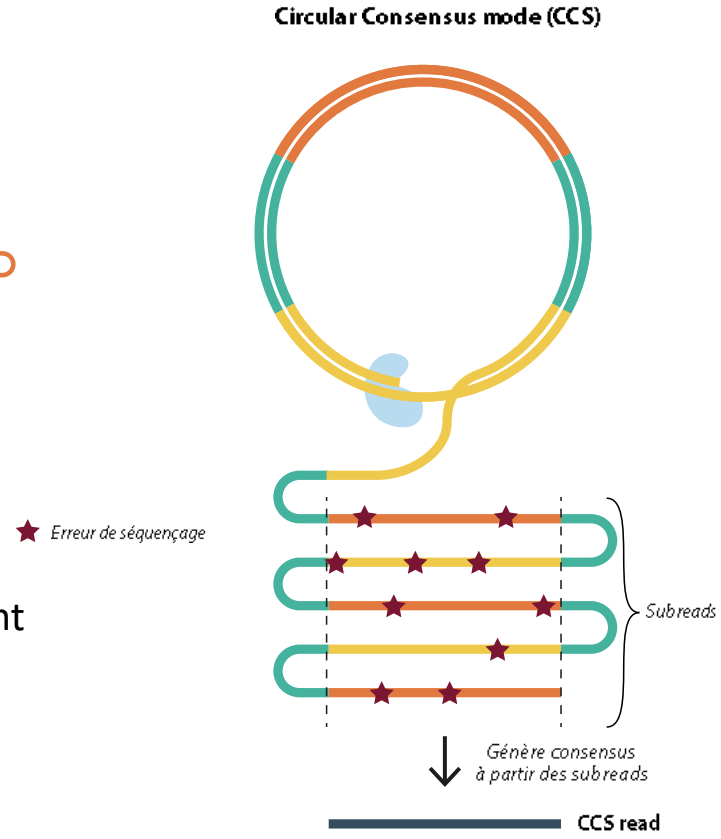
→ Fragments long mais qualité de séquence **moyenne** (~85% identité)



Circular **C**onsensus **S**equence (CCS ou Hifi)

Obtention d'une séquence **consensus** pour chaque fragment, **plusieurs lectures** par fragment

→ Fragments plus courts (10kb) mais de **meilleure** qualité (>99.9% identité)



2. Génomes complets et assemblage *de novo*

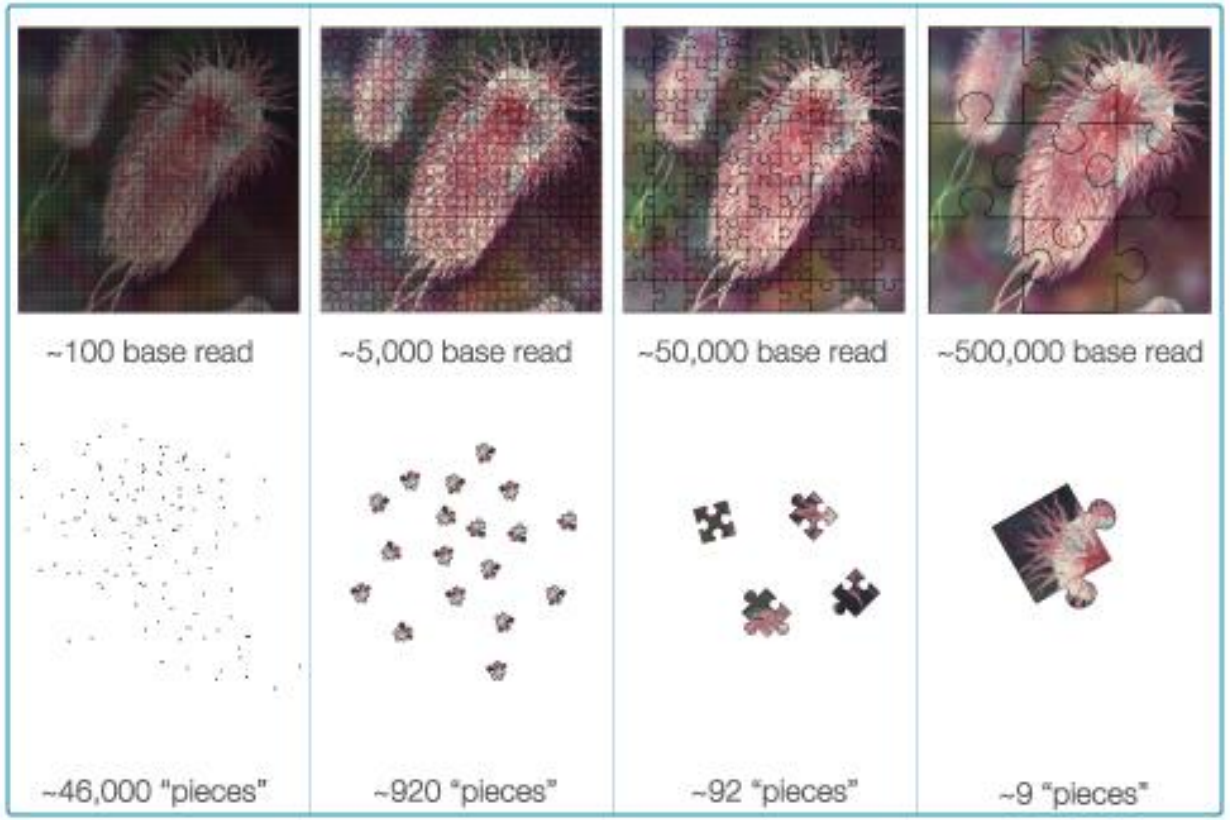
GenoScreen
Innovative Genomics

Assemblage *de novo*

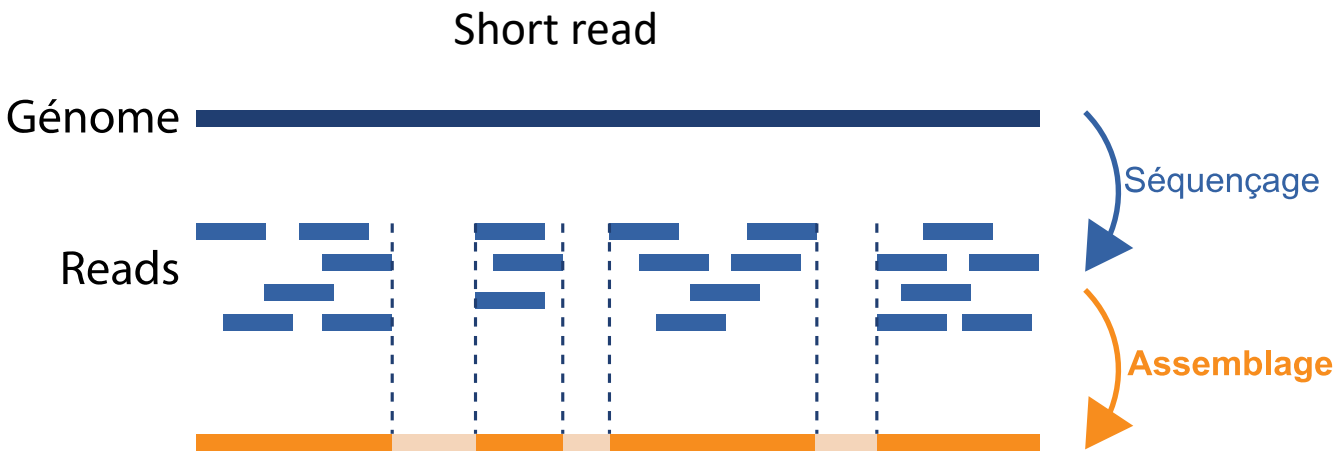
Reconstruire le génome uniquement à partir des reads, **sans** se baser sur une **référence**.

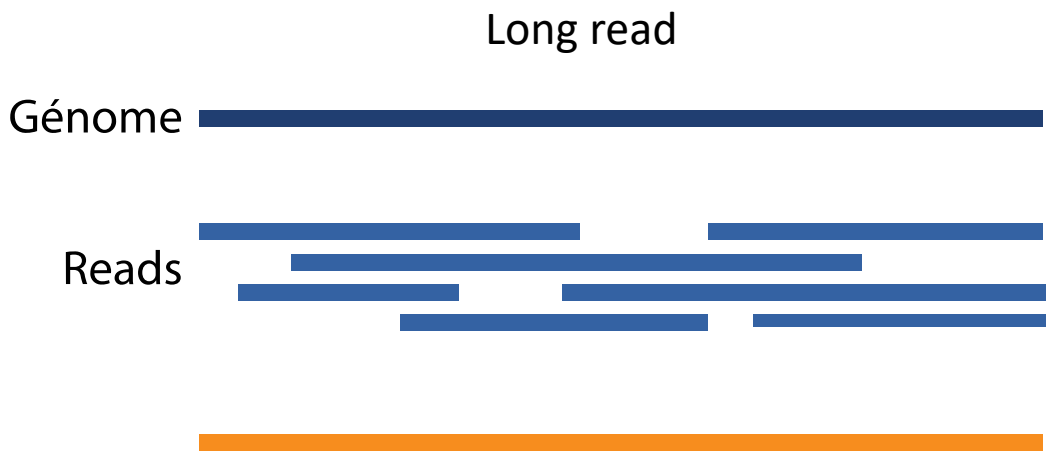
→ Vision la **plus complète** et **moins biaisée** d'un génome

➔ Long read = permet **d'assembler** plus facilement des génomes **complets**



Assemblage *de novo*





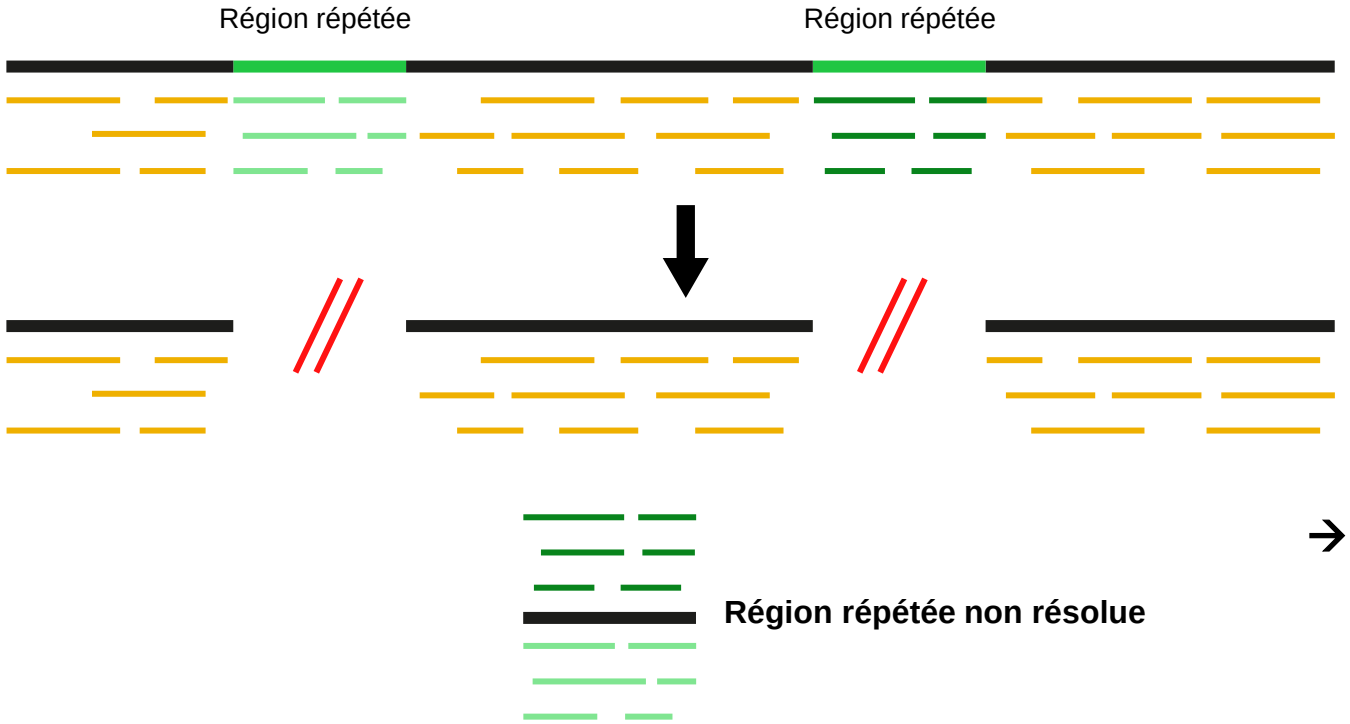
→ Information **manquante**, génome **fragmenté**, apparition de **gaps** dans l'assemblage

→ Vision complète de **l'architecture** du génome, de sa composition **fonctionnelle** et **structurelle**

➔ **Pourquoi les assemblages avec des short reads sont-ils plus fragmentés ?**

Short read et assemblage

Impact des répétitions lors de l'assemblage



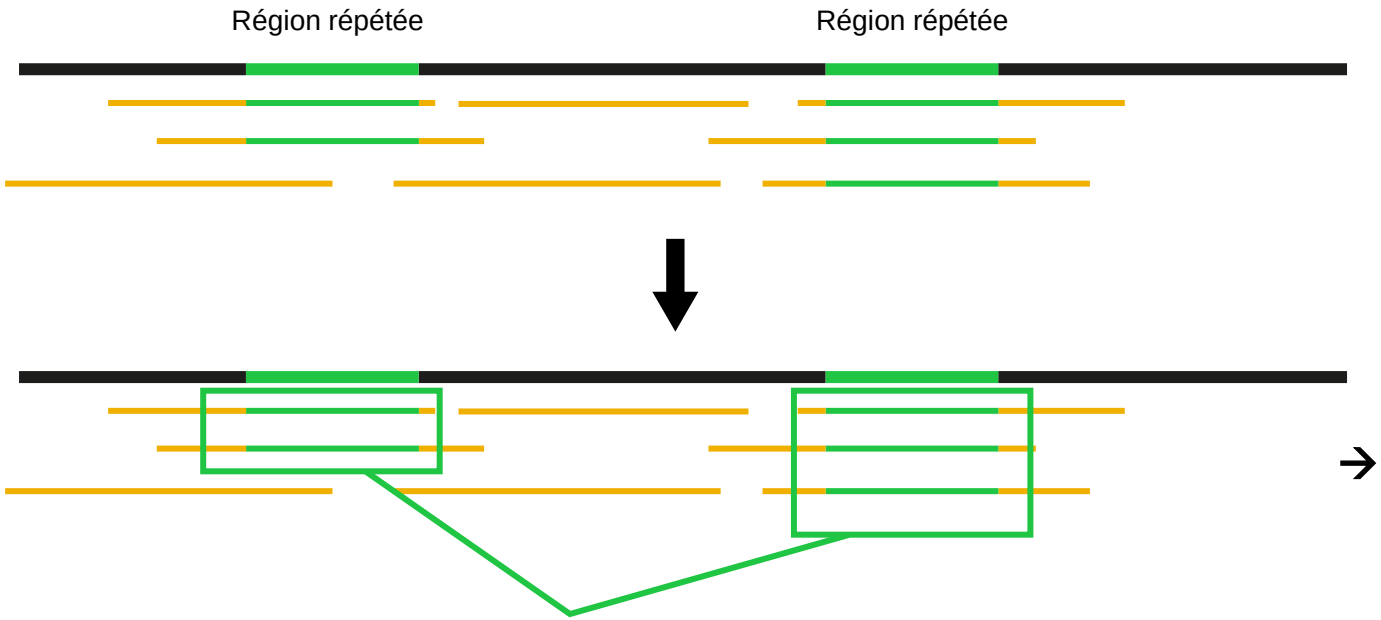
→ **4 contigs**

Impossible d'ancrer les répétitions dans leur contexte en amont et en aval

→ **Création de points de cassure de l'assemblage**

Long read et assemblage

Utilité des long reads



→ **1 contig**

Répétions ancrées dans leur contexte génomique en amont et en aval

➔ Répétition résolues

Condition *sine qua non* pour résoudre une région répétée :

Longueur des reads > longueur de la région répétée

3. Obtention de génome de référence : Cas du génome humain

GenoScreen
Innovative Genomics

Vers des génomes complets : 1ère génération

Human Genome Project (1990-2003)

But : obtenir un génome de référence pour l'Homme (méthode Sanger)

- 3 Milliards d'euros
- 13 ans
- Consortium international

Révolution pour la génomique

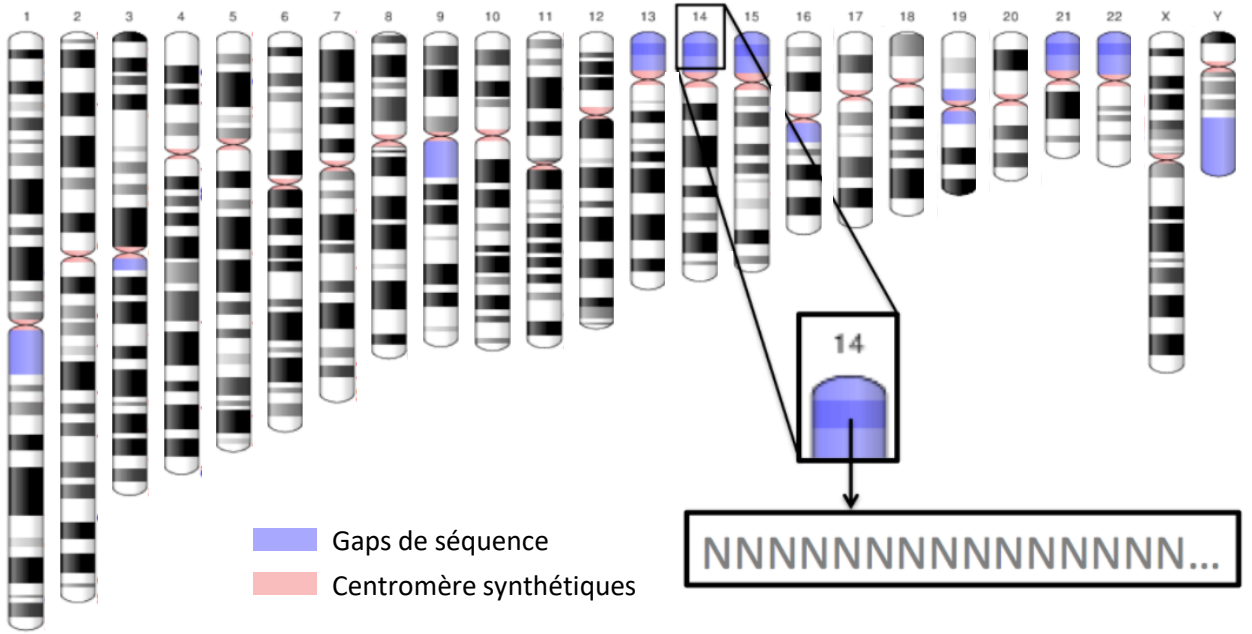
Bonne couverture des séquences géniques mais moins le reste
→ Mais… Enormément de **gaps**, « seulement » **2,5 Gb** déterminées

Vers des génomes complets : 2ème Génération

Genome Reference Consortium

Travaille depuis 20 ans à l'amélioration de la version initiale

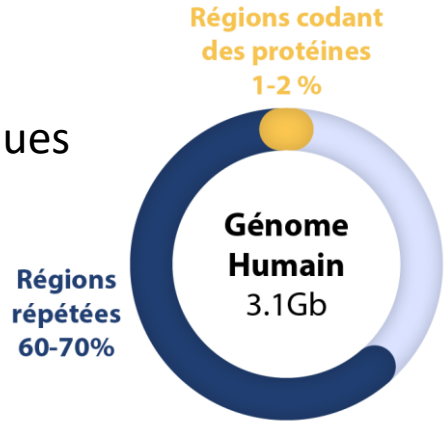
Désormais en version GRCh38.p14



Toujours de grandes régions manquantes :

- Centromères (séquences synthétiques)
- Régions péricentromériques
- Bras p chromosomes acrocentriques

→ Correspondent à des **régions répétées**



http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/

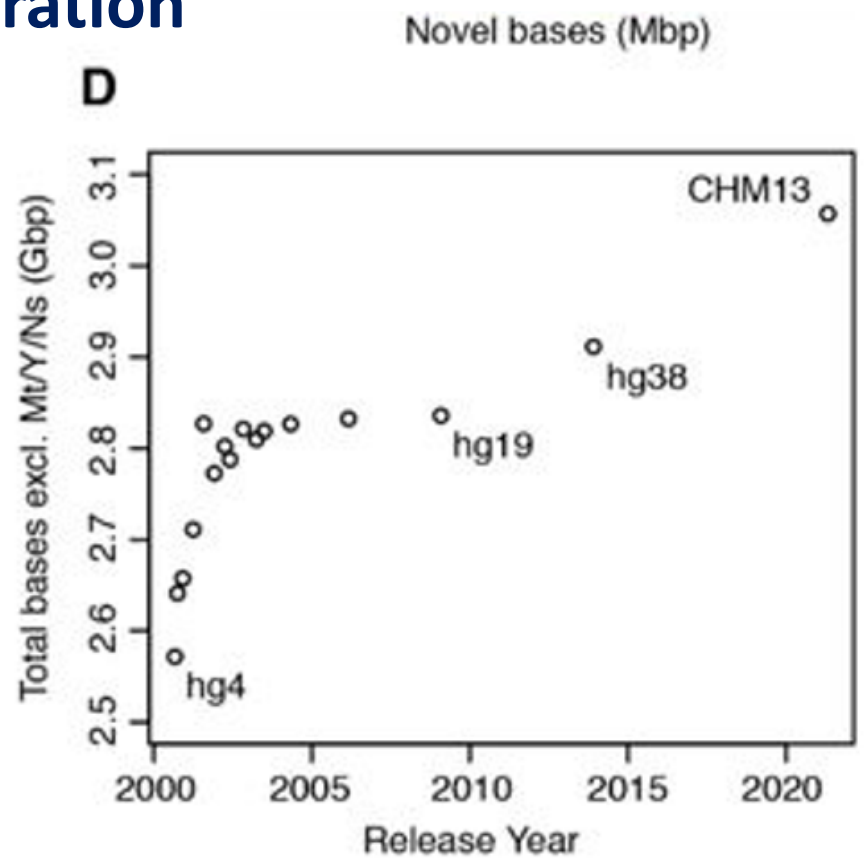
Vers des génomes complets : 3ème génération

Telomere-to-Telomere Consortium

But : obtenir la séquence **complète** et sans gap du génome humain

Combinaison de plusieurs techniques

- Ultra long read nanopore
- PacBio Hifi (CCS)
- + autres méthodes (Hi-C, mapping optique)



Grâce aux technologies long read (et beaucoup de validations manuelles et expérimentales) :

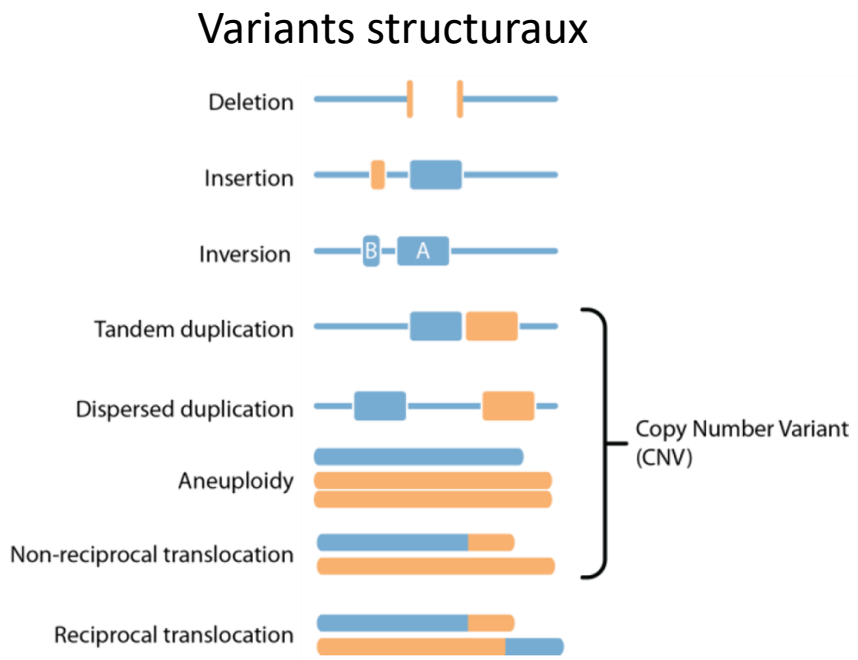
- Ajout de près de **200Mb**
- Résolution des gaps du GRCh38 : **Centromères**, **Chr acrocentriques**, etc.
- Manque encore **chromosome Y**…(lignée cellulaire utilisée homozygote XX)

Nurk et al., bioRxiv, 2021

Enjeux en santé humaine

Détection de nouvelles associations variant / maladie

- Répétome =/= junk DNA (ADN poubelle)
  - Variations dans des régions répétées impliquées dans cancer, diabetes, Trouble du spectre autistique, maladies mentale (schizophrénie et dépression)
- Variants structuraux
  - Possible avec short read mais plus robuste avec long-read
  - Détection de variants complexes



Enrichissement du génome de référence pour représenter l'espèce entière (pangénome)

- Diagnostic et prognostic égal pour toutes les populations

4. Autres exemples d’applications

GenoScreen
Innovative Genomics

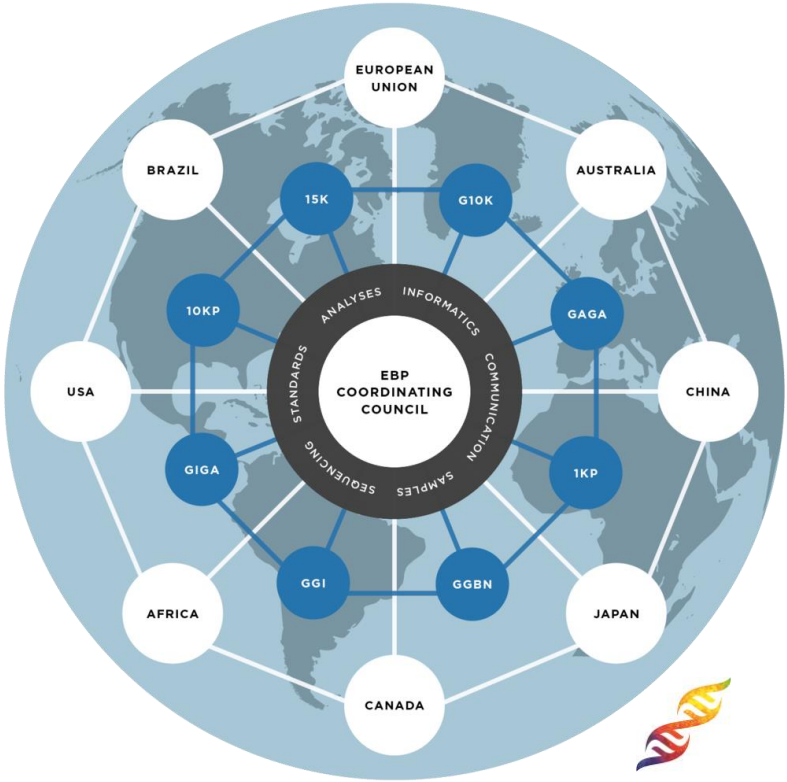
Vers une vue complète et non biaisée de la diversité naturelle

Earth BioGenome Project

Obtention de génomes complets pour 1,5 millions d'espèces eucaryotes sur une période de 10 ans

Buts :

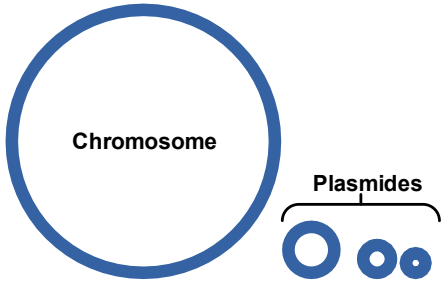
- Compréhension globale des mécanismes d'évolution
- Conservation, protection et restauration de la biodiversité
  - Compréhension du fonctionnement des écosystèmes dans leur globalité
- Apports à l'amélioration de la société et le bien-être humain
  - Création de nouveaux biomatériaux, biofuels synthétiques
  - Nouvelles molécules thérapeutiques



Applications en microbiologie

Facilité d'obtention de génomes complets circularisés

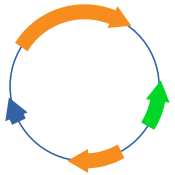
- Génomes de référence
- Caractérisation de souches d'intérêt industriel



Après assemblage, 1 contig = 1 réplicon

Vérification de plasmides

Possibilité d'obtenir la séquence de plasmides rapidement (<2h) avec un coût maîtrisé



Métagénomique ciblée

Séquençage de l'ADN ribosomique 16S dans son intégralité (1,6kb) = meilleure résolution taxonomique

Métagénomique shotgun

Obtention de génomes complets à partir d'échantillons complexes !

Trancriptomique

Séquençage de transcrits complets

- Assemblage *de novo* de transcriptome
- Découverte de nouvelles isoformes
- Analyse d'expression d'isoformes (différentielle ou non)
- Détection de gènes de fusion

Epigénétique

Comment détecter des bases modifiées ?

3ème génération séquence sur une molécule native

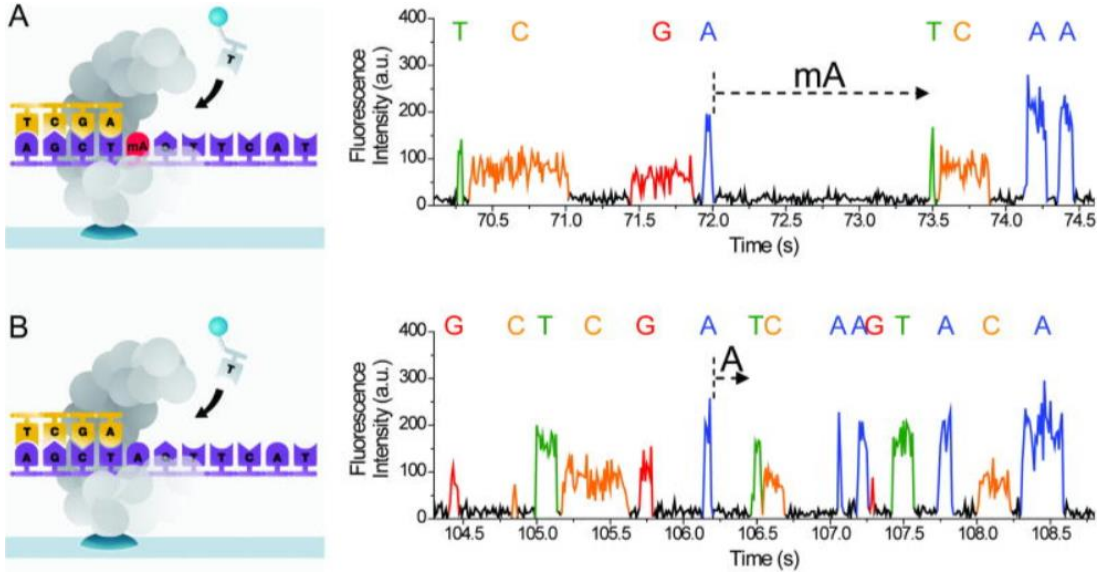
→ Pas d'amplification PCR, **rétention** de toute l'information **épigénétique** sans préparation particulière de l'échantillon

Base modifiée = Encombrement stérique plus important

- PacBio → Temps plus long pour incorporer la base suivante
- ONT → Signature électrique différente au passage d'une base modifiée

  + capacité de lecture modification ARN ! (lecture directe ARN)

→ Utilité également pour le **phasing** de génomes (ploïdie >1n) afin de retrouver des **haplotypes** complets



Fusberg *et al.*, Nature Methods, 2010

Conclusion

- Permet d'améliorer des génomes incomplets (draft)
  - Augmentation de la contiguité (moins de contigs)
  - Meilleure compréhension de la structure des génomes
- Permet l'obtention de génomes complets et contigus pour n'importe quel organisme
  → Parfois besoin de corriger les séquences par un « polishing » avec short reads
- Lecture de molécules natives permet détection de bases modifiées
- Technologies toujours en cours d'amélioration, arrivée prochaine de « nouveaux » acteurs (Illumina Infinity)
- Également possibilité de faire du séquençage avec des long read synthétiques (LoopSeq, TellSeq, etc.) basé sur des technologies NGS.

Merci de votre attention

GenoScreen
Innovative Genomics

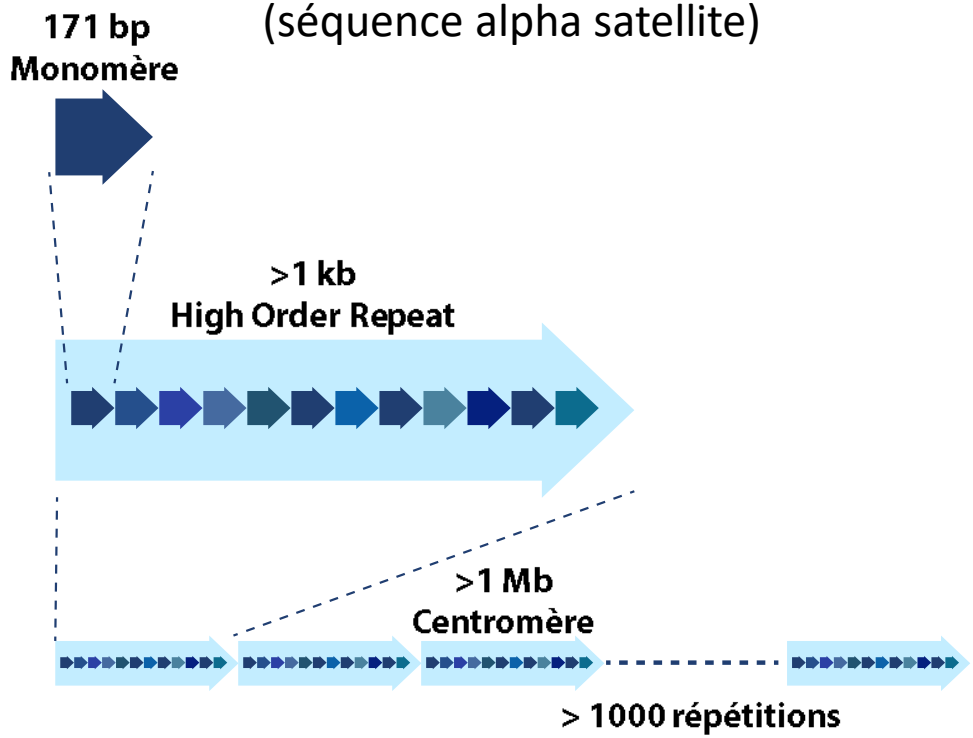
Siège social
1 rue du Pr. Calmette
59000 Lille
FRANCE

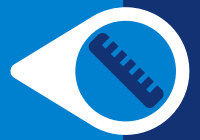
03 62 26 37 77
contact@genoscreen.fr
www.genoscreen.fr

Pourquoi ces gaps « résistent-ils » au séquençage ?

Régions répétées du génome

Structure d'un centromère
(séquence alpha satellite)



Oxford Nanopore Technologies (ONT)		Pacific Biosciences (PacBio)
	Taille des lectures	
Théoriquement illimitée		50-100 kb CLR / 10-15 kb CCS
	Fidélité	
Erreurs systématiques dans homopolymères = couverture ne résout pas le problème		Erreur aléatoire : permet avec CCS (HiFi) d'obtenir reads quasi parfait
	Lecture directe	
Bases modifiées (méthylation) lecture directe ARN		Bases modifiées (ADN seulement)
	Matériel de départ	
Longueur et quantité des lectures dépendants de la qualité du matériel extrait (ADN/ARN)		Quantité et qualité du matériel génétique très importants
	Analyse en directe	
Basecall en temps réél «Adaptive sampling»		Film en direct mais analyse a posteriori
	Coût	
Coût avantageux pour des projets de petite et moyenne taille		Si CCS, permet de se passer de Illumina pour corriger

Métagénomique ciblée

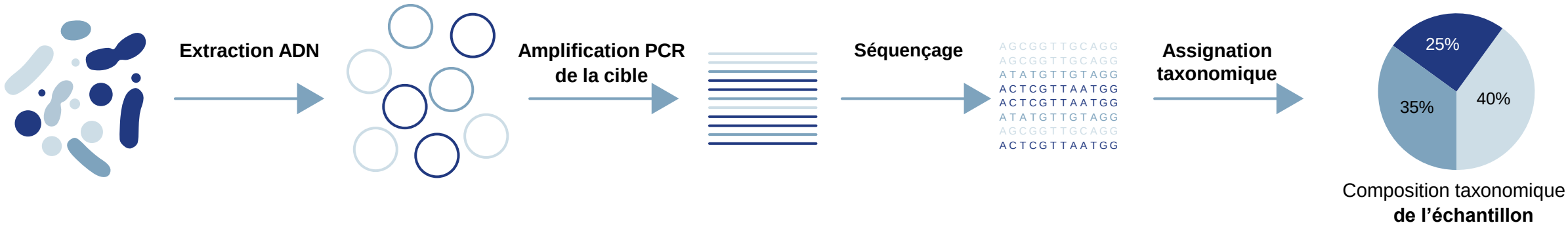
Qui est là?

Autre nom : Metabarcoding

But :

Informe sur la **structure** et la **composition** de la population

Principe :



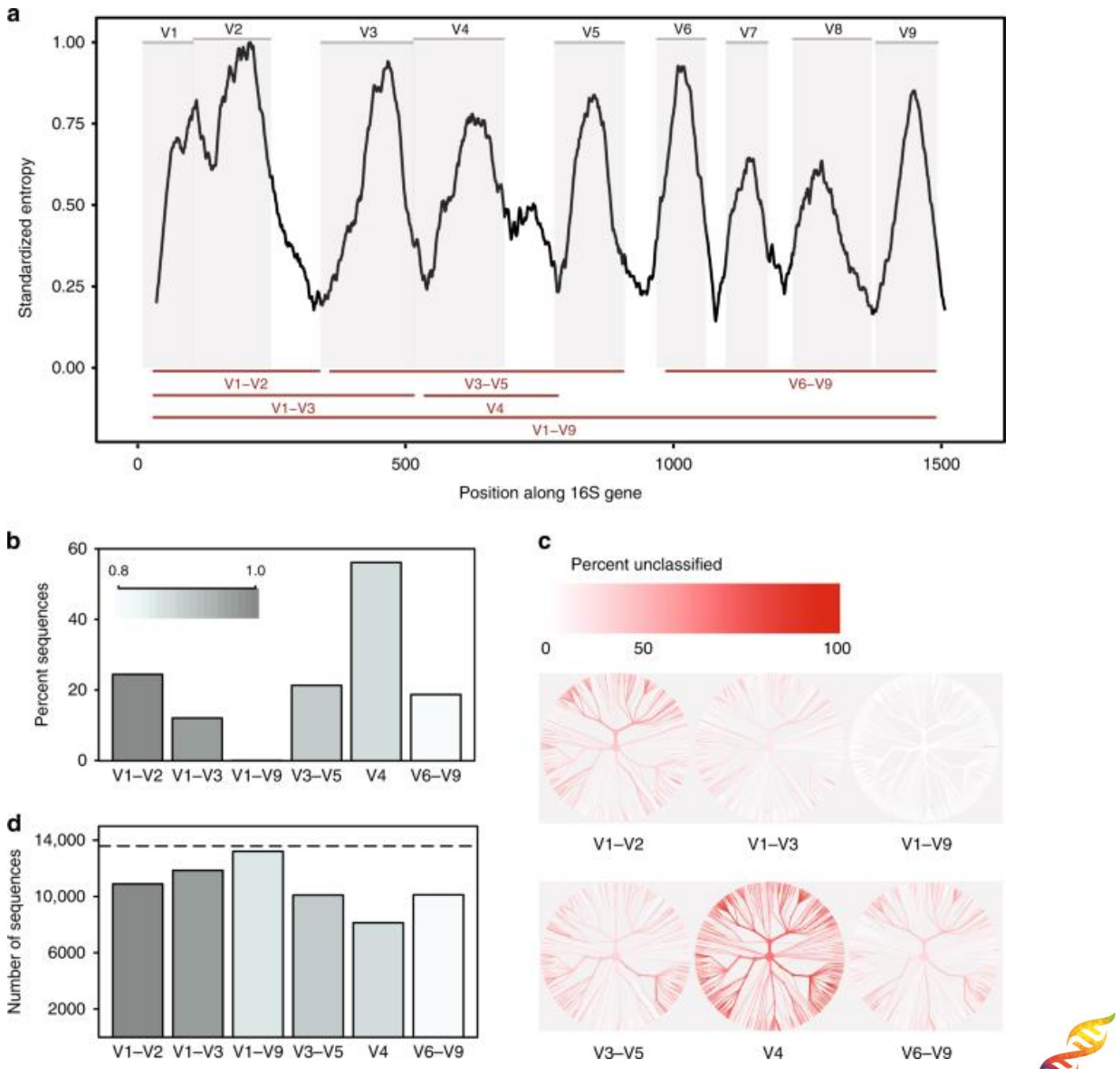
Composition taxonomique **de l'échantillon**

Métagénomique ciblée

Cible ARN ribosomique 16S (18S chez eucaryotes)
Taille totale 1 600 pb

Séquençage de 2nde génération :
Séquence qu'une partie du 16S (max 300 pb)
= perte d'information
→ Résolution possible de la taxonomie jusqu'au genre

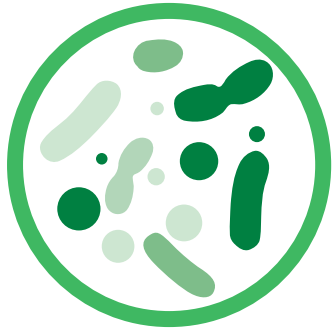
Séquençage de 3ème génération :
séquence complète du 16 (voir même plus)
= conservation de toute l'information
→ Résolution taxonomique jusqu'à l'espèce



Johnson *et al.*, Nature Communications, 2019

Métagénomique shotgun

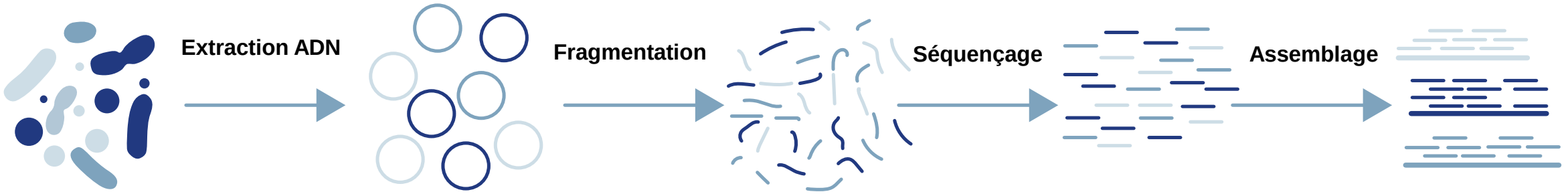
Qui **peut faire** quoi ?



Buts :

- Informe sur le **potentiel fonctionnel** et **physiologique** des organismes en présence (Gènes / Voies métaboliques)

Principe :

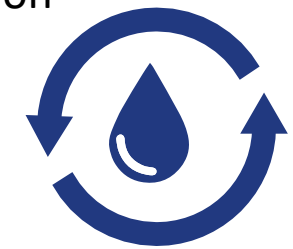


Métagénomique shotgun

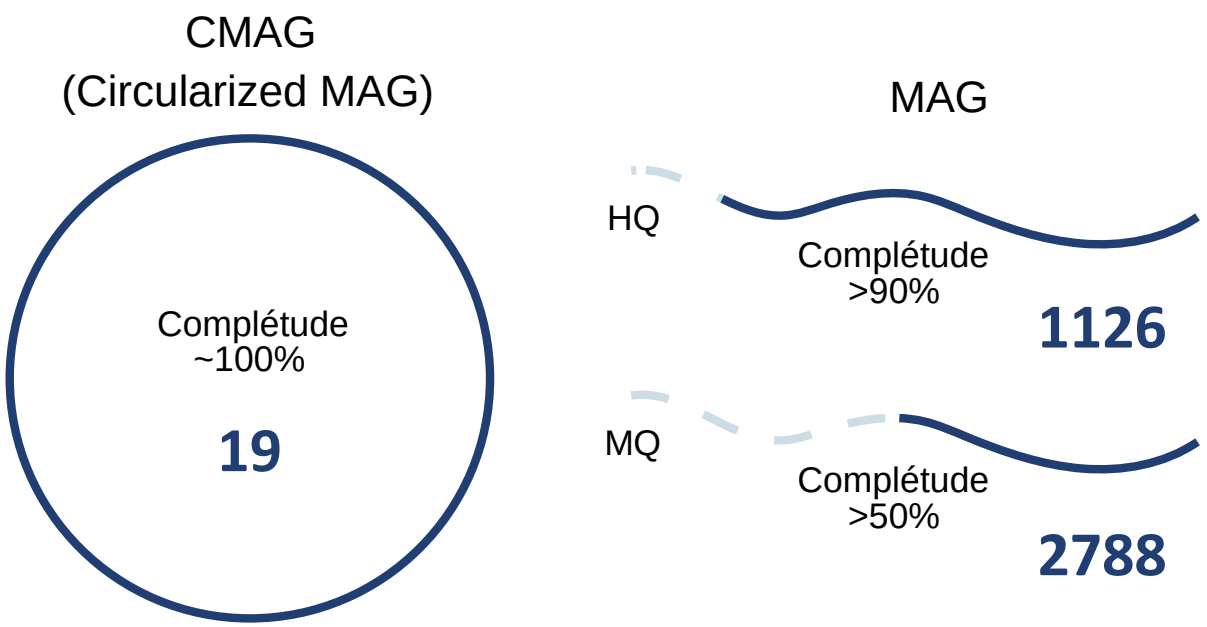
MAGs → Utiles pour évaluer le **potentiel fonctionnel** des communautés

Metagenome **A**ssembled **G**enomes

23 sites de prélèvement de stations d'épuration



Grâce au long read :
3733 MAGs assemblés



Singleton *et al.*, Nature Communications, 2021