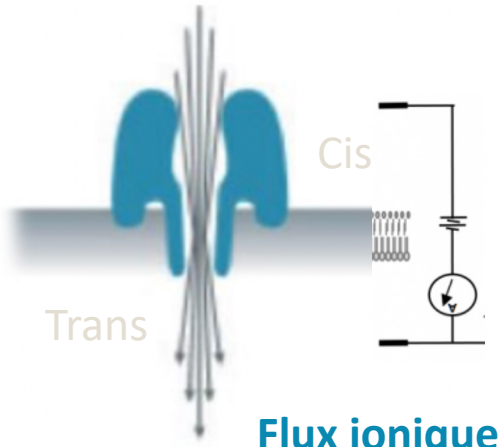
Séquençage Nanopore
&
Métabarcoding

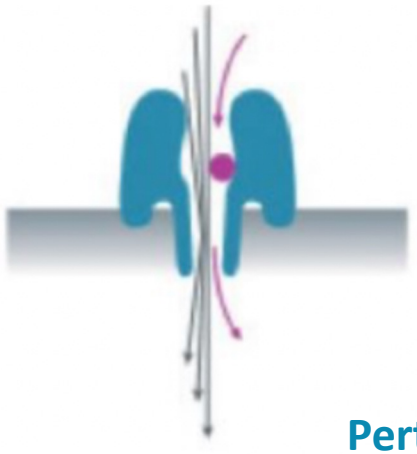
Fabrice Armougom, IRD
Mediterranean Institute of Oceanography
fabrice.armougom@mio.osupytheas.fr

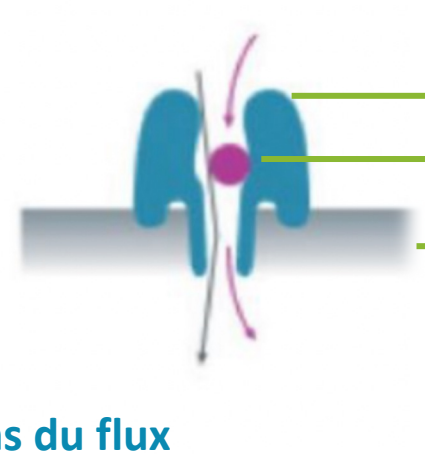
réseau BIOinformaTIque en provenCe

Local MIO, Mars 2026

Principe Général de Détection par Nanopore







Nanopore

Molécule

Membrane isolante

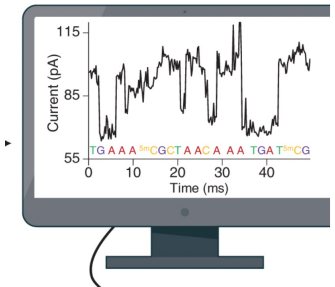
Flux ionique

Perturbations du flux

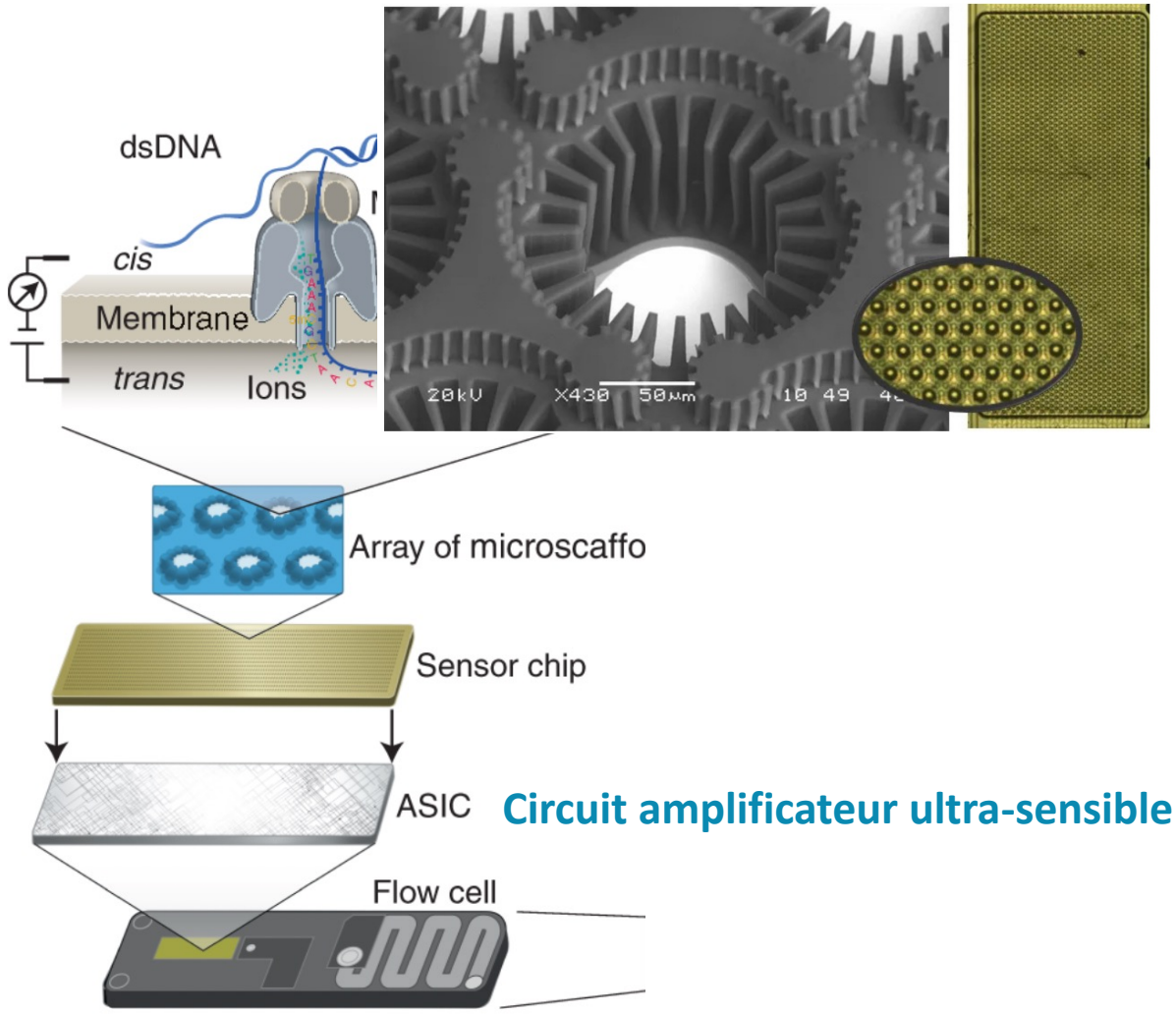
Courant référence

Empreintes électriques

Décodage : Traitement du Signal



Structure d’une Flow Cell Nanopore

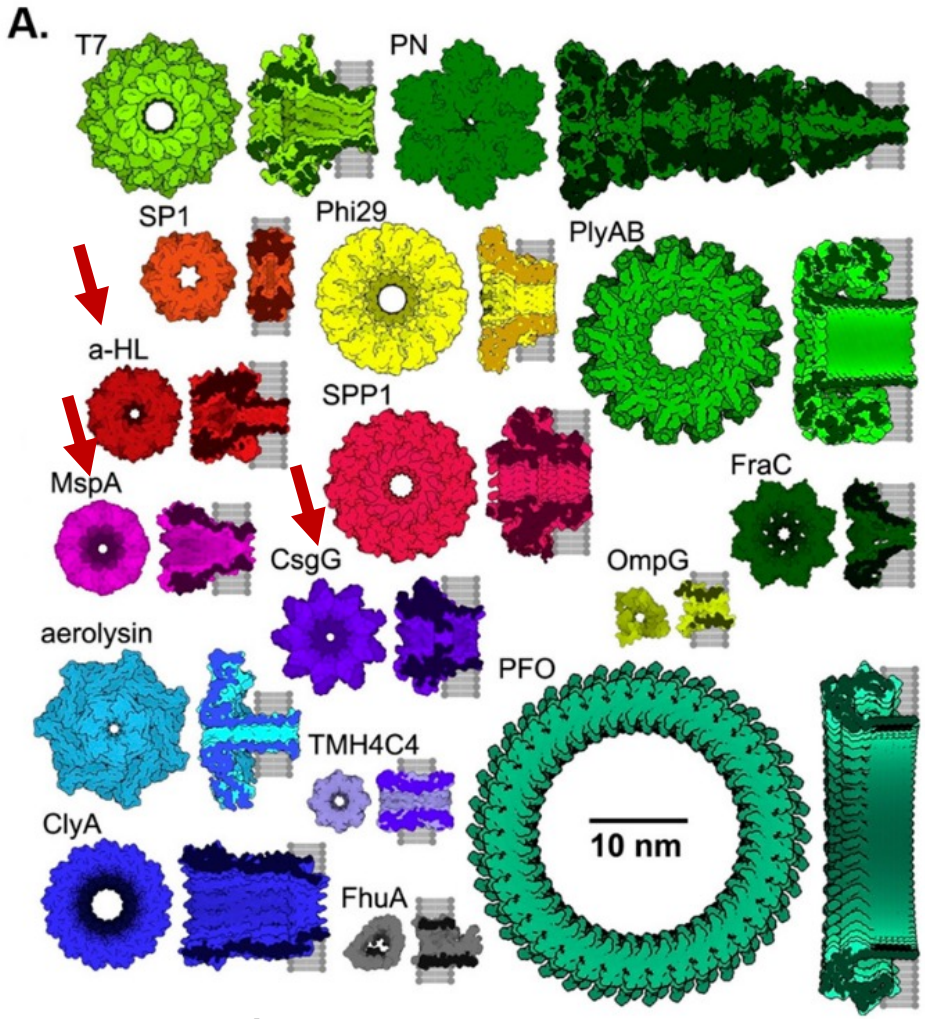


Microstructures avec des électrodes (sensor chip)

Nanopore inséré dans une membrane polymérique

Circuit amplificateur ultra-sensible (mesure/contrôle)

Les Principaux types de Nanopores « Naturels »



Mayer SF *et al.*, 2022

Grande variété, longueur, diamètre du pore

Biocapteurs

Biomédical, Environnement, Nanotechnologies

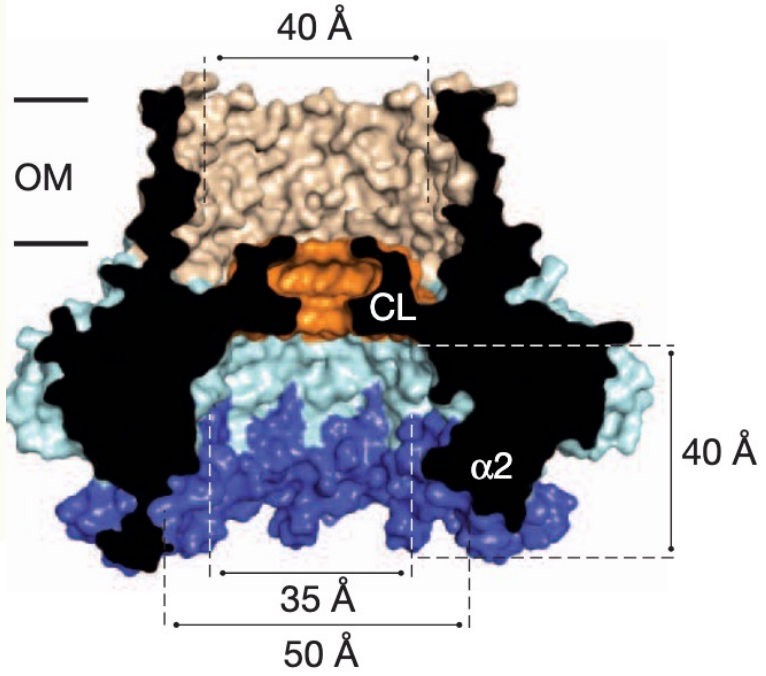
- **Détection de petites molécules : LamB/FhuA (sucre, métabolites)**
- **Détection de peptides/protéines : ClyA**
- **Nanomoteur ADN : Phi29**
- **Valves, contrôle libération de molécules**

Séquençage Nanopore : Le pore CsgG (Curli Specific Gene G)

Goyal *et al.*, *Nature*, 2014

Pore CsgG : Canal de sécrétion, passage pour CsgA/B (fibres curli = biofilm)

- **Entrée Ø ~ 4 nm, Etranglement Ø ~ 1,5 nm**
- **Longueur du canal pore : ~ 9 nm**
- **ADN simple brin Ø ~1 nm**

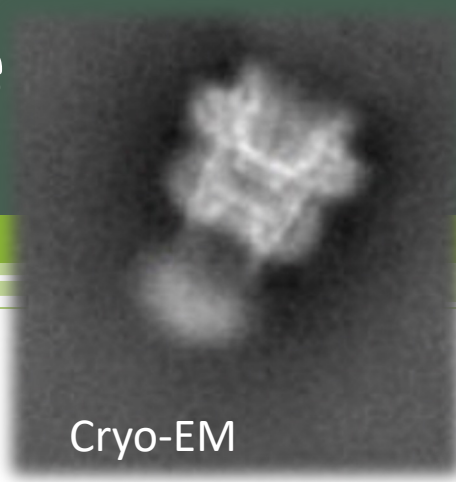


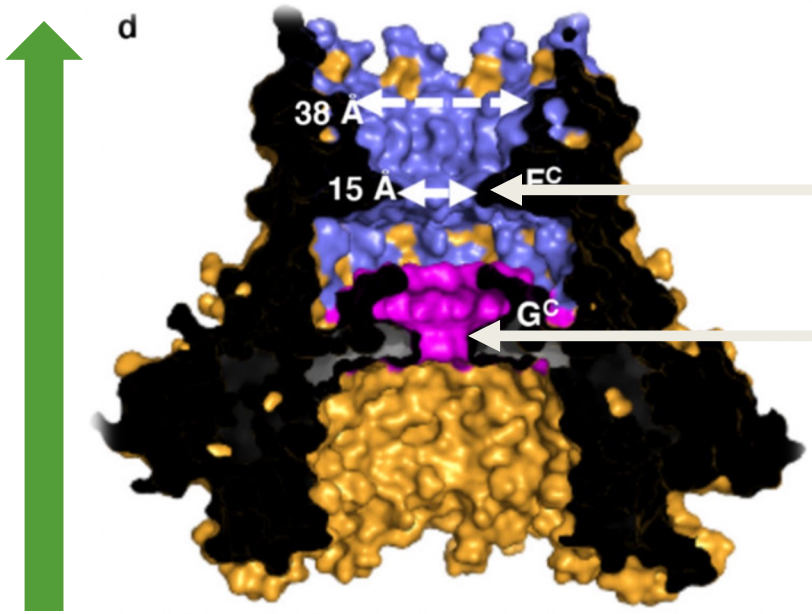
Pourquoi CsgG?

Stabilité, Ø pore adaptée ssDNA, flexibilité ingénierie, production, …

Le pore CsgG-CsgF : Double Constriction et Stabilité

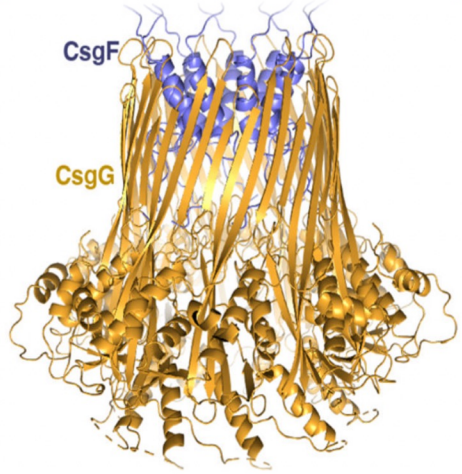
Sander E. Van der Verren *et al*., 2020, Punam Rattu et al., 2021





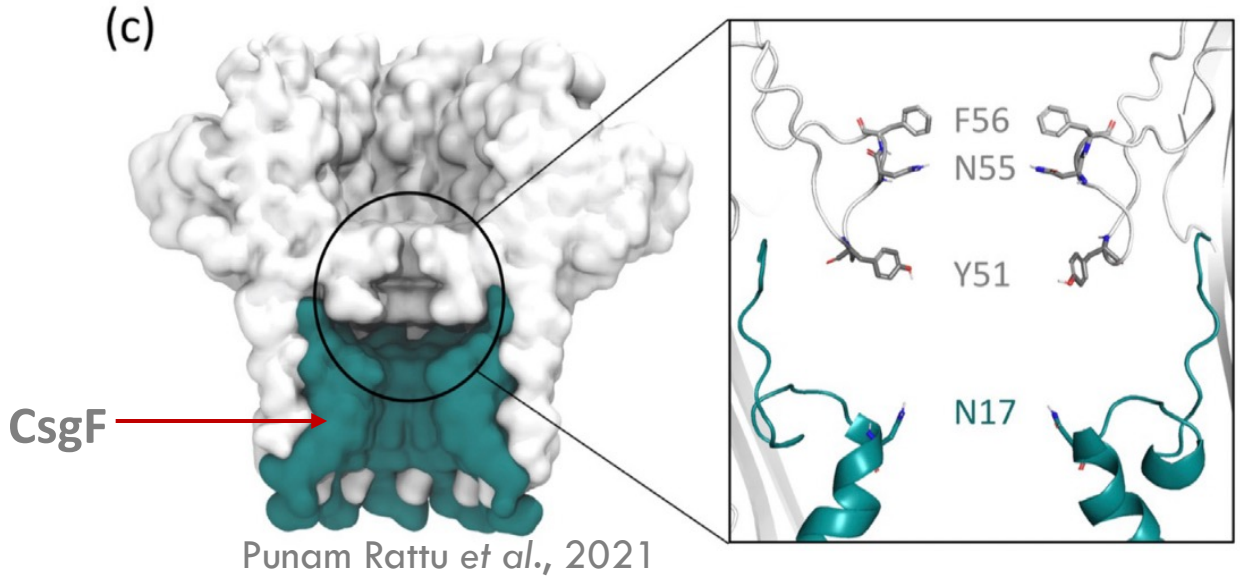
2eme Constriction liée à CsgF : 1,5 nm

1ere Constriction native CsgG : 1,7 nm



C’est au niveau des « étranglements »
que l’ADN est réellement « lu » par le pore

Zones d'Etranglements et Interactions



Punam Rattu *et al.*, 2021

Interactions transitoires

- π-stacking
- Electrostatiques
- Liaisons hydrogènes

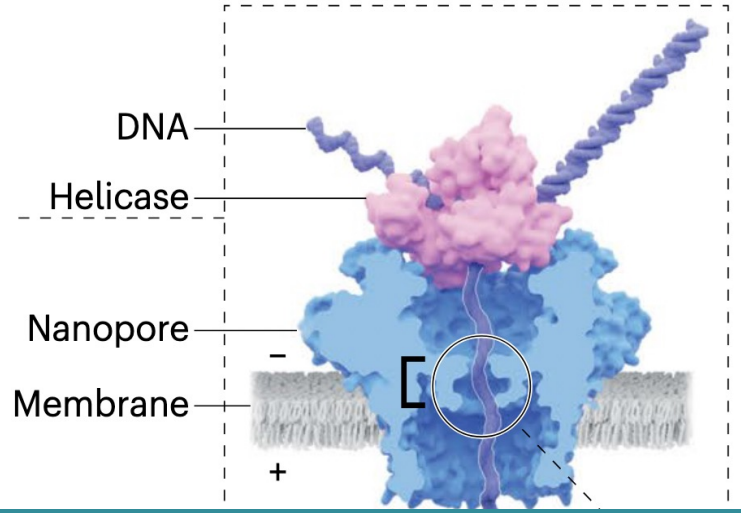
Les AA des zones d'étranglement interagissent avec les bases ADN et modulent le flux ionique, donnant au signal sa spécificité

→ Empreinte électrique = "**squiggle**"

→ Zones étranglement est un **lecteur de K-mers** (4-6 bases)

Pore CsgG vs CsgG-CsgF

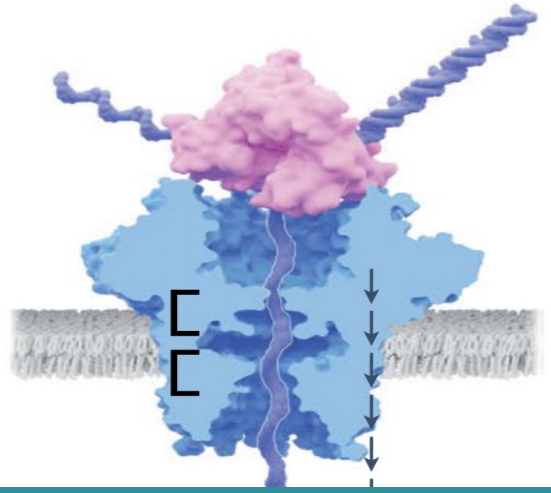
CsgG = Flow Cell R9



Simple lecteur

- Précision 94-97% (Q12-15)
- Débit plus Rapide
- Meilleur rendement

CsgG-CsgF = Flow Cell R10

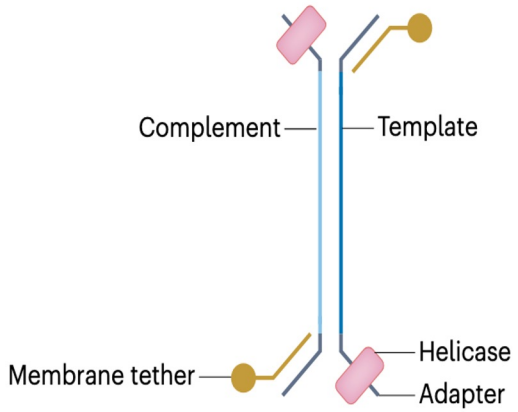


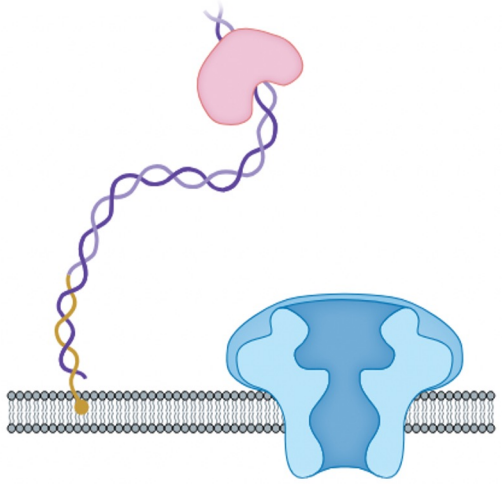
Double lecteur

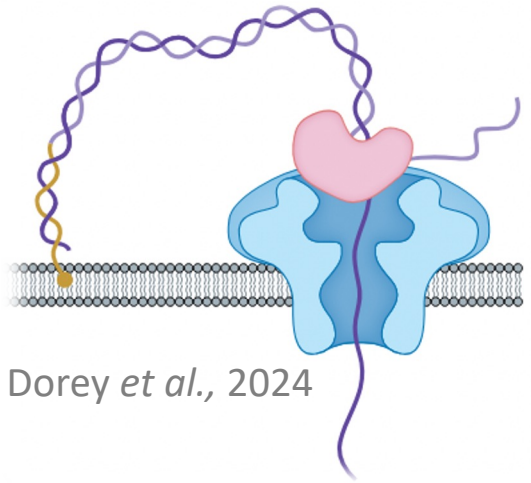
- Précision 99,9% (Q30)
- Réalité (Q25)
- Résolution homopolymères ++
- Plus complexe → algorithmes adaptés

Protéine Motrice du Séquençage Nanopore : Polymérase à l'Hélicase

Utilise des Hélicases ATP-dépendantes :
- **Frein moléculaire, Dérouleur (ssDNA), Régularité**
- **Préfixées aux adaptateurs dans les KITs!**



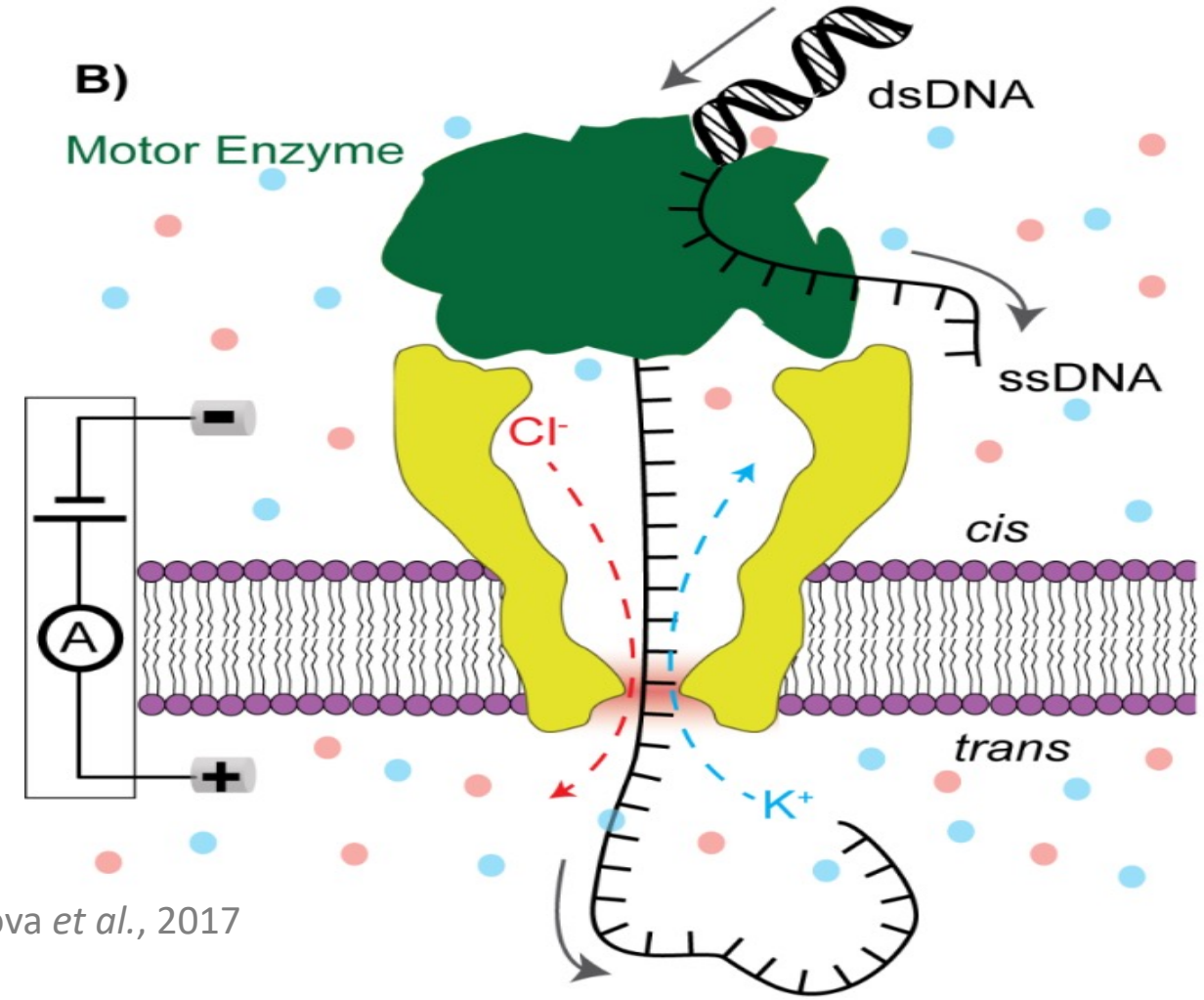




A. Dorey *et al.*, 2024

Sans hélicase : ssDNA file à ~10^6 bases/s = signal illisible
Avec hélicase : vitesse ralentie à ~400 bases/s = signaux exploitables

Séquençage Nanopore : Synthèse



Ian C. Nova *et al.*, 2017

Ancrage ADN
Hélicase (ssDNA)
Diff potentiel → Courant ionique
Déplacement K^+/Cl^-
signal unique : Squiggle

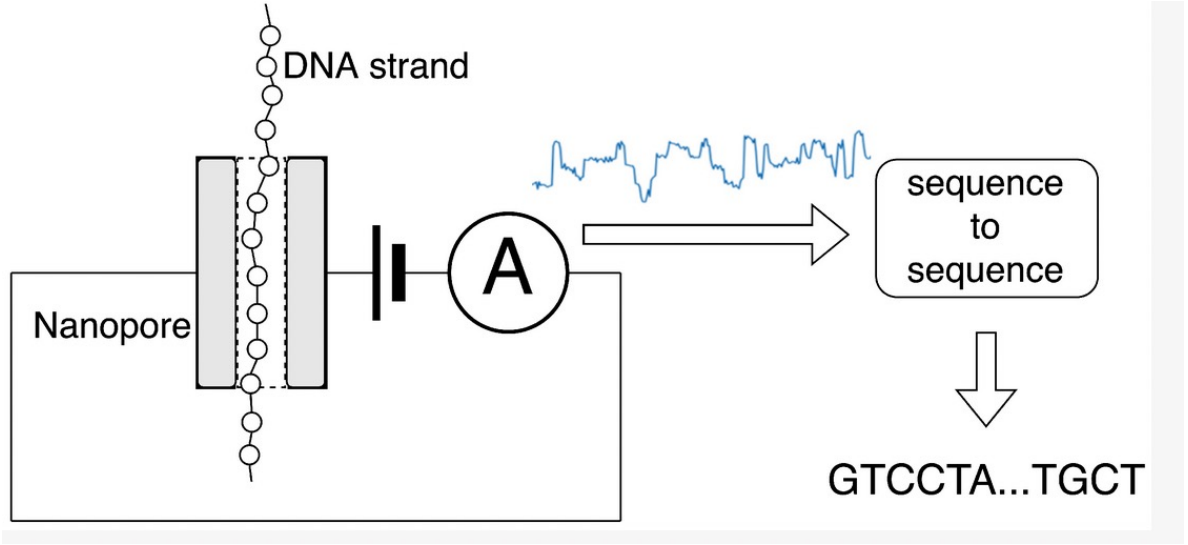
Taux Erreur nanopore

Technologie	Type d'erreur dominant	Homopolymères	Q-score moyen (approx.)
Illumina	substitutions	OK	Q30
Nanopore	**insertions / délétions**	difficile	Q10–Q20 (jusqu'à Q20+ récent)
PacBio HiFi	erreurs aléatoires (insertions)	OK	Q30–Q40+

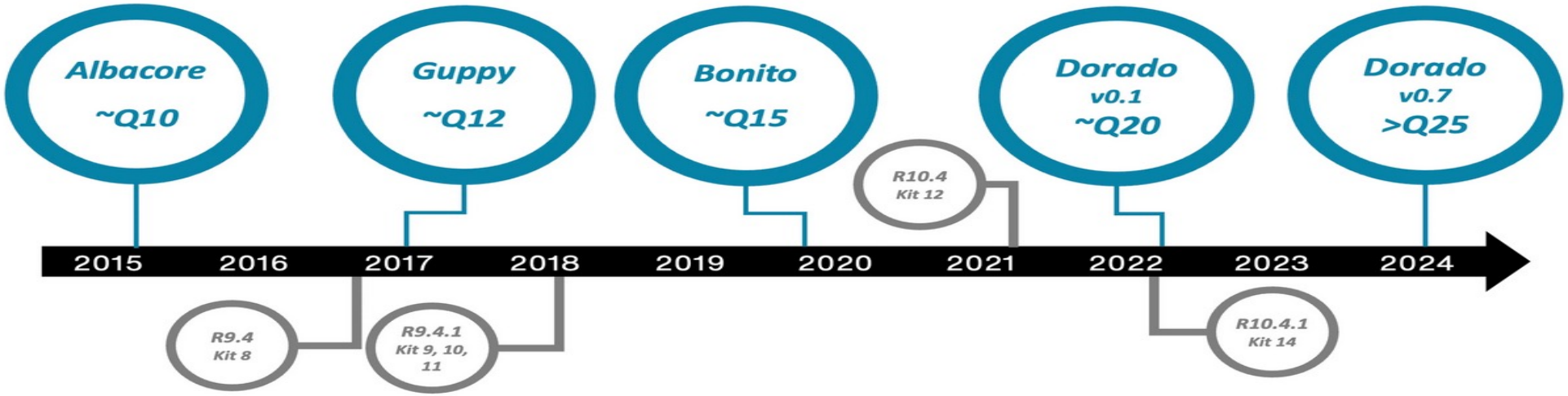
Nanopore : Q-scores plus faibles et une **proportion élevée d'erreurs d'insertion et de délétion** → Améliorations récentes des algorithmes de basecalling.

Basecalling : Conversion du Signal Electrique Brut en Séquence de Bases

grâce à des algorithmes de deep learning toujours plus performants



Evolution des Basecallers!



Séquentiel (RNN) = LENT!! Décodage Flip-Flop	motifs locaux (CNN) GPU/rapide positions voisines /décodage (CRF)	Motifs locaux (CNN) GPU/rapide vision globale/décodage (Transformer)

En 10 ans, on est passé de ~90 % à >99 % (> Q25) de précision
RNN → CNN/CRF → CNN/Transformer

Basecalling & Apprentissage

Apprentissage : Alignement des **signaux bruts** sur des **séquences de référence connues**

But : **associer motifs électriques aux bases** (A, C, G, T) grâce à des millions d'exemples, validés ensuite sur des jeux indépendants

3 Modes pour le basecalling : de la Vitesse à la Précision

FAST : petit réseau, peu de calculs = rapide mais moins précis (Adaptative Sample)
HAC : taille intermédiaire = compromis (Défaut, temps réel)
SUP : gros réseau, plus précis mais nécessite plus de temps/ressources (Post séquençage)!!

Formats des sorties : POD5, Fastq et Rapports

Signal Brut

- FAST5 : Signal brut + métadonnées techniques (Flowcell ID, Chimie, paramètres séquençage, état des pores, info samples etc)

→format LOURD!!

- POD5 : optimisé pour la vitesse d'écriture et le big data (2023)

Format de Séquences : Fastq

Format alignés (optionnels) : BAM, CRAM

Rapports (HTML, CSV)

Graphiques et stats sur le run (pores, output cumulatif, Q score, vitesse)

Fichier spécifique

Sequencing_summary.txt (pycoQC, Nanoplot …)

Digression : Adaptative sampling : Séquençage Ciblé en Temps Réel

Mode Dépletion

Squiggle matching = signal brut

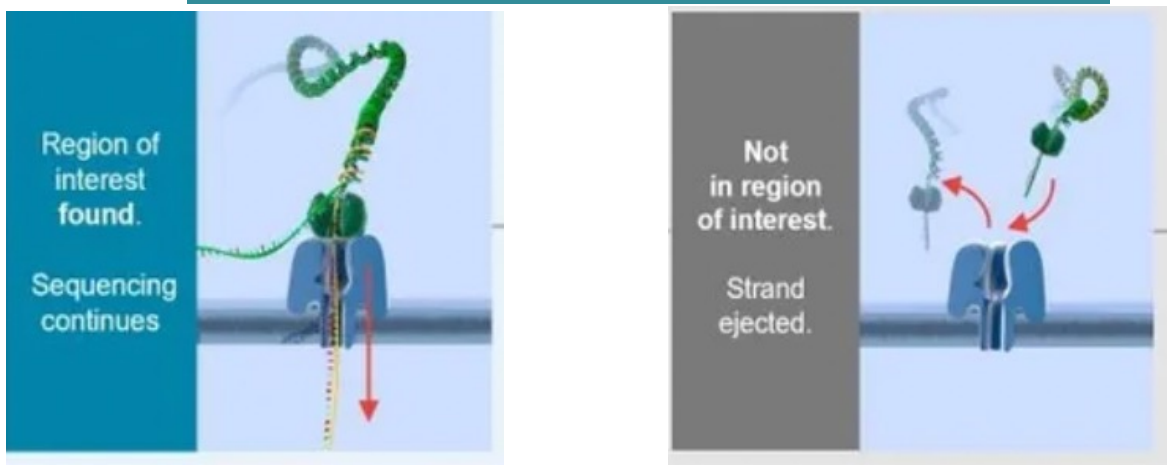
- Rapide
- Eliminer séquences hôte/…

Mode sample select

Demultiplexing-aware adaptive sampling

Mode Enrichissement

Basecalling précoce = séquence



Enrichissement ciblé d'un gène (fonctionnel) d'intérêt

Limite : Ne lit que les **400 premières bases**

Base de données

Vers des modèles plus compacts, plus intégrés

MinION MK1D

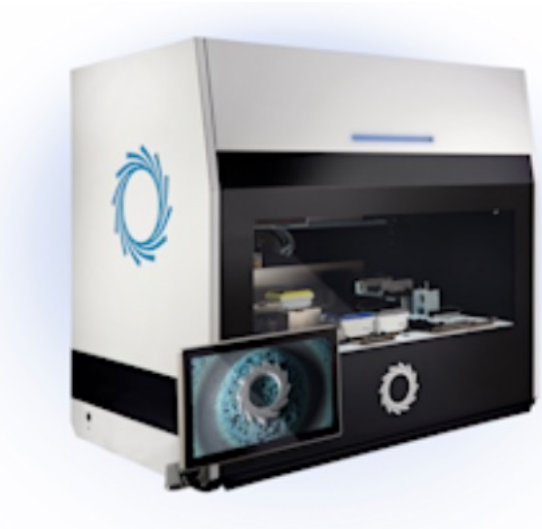
3 K€

60 K€

GridION

80 K€

PromethION 2 Integrated (P2i)



ElysION 250 K€

Flonge*

PromethION 24 (P24) 400 K€

PromethION 2 solo (P2 solo)

26 K€

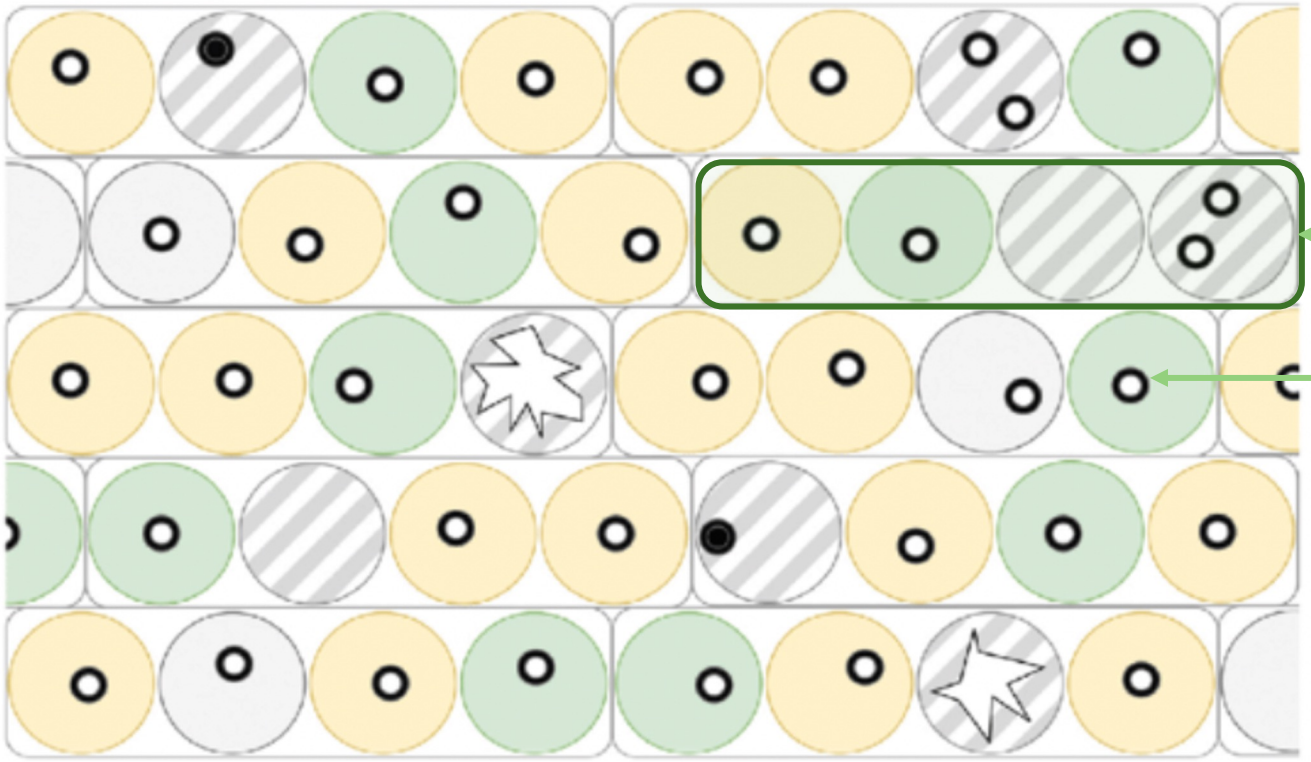
Gamme ONT : Débit

Appareil	Nombre de FlowCell	Nb pores	Canaux actifs	Débit/Flowcell
Flongle	1	126	~126	1–2 Gb (up to 3.3 Gb)
MinION Mk1D	1	2048	512	10–30 Gb (up to 38 Gb)
GridION	Up to 5	2048	512	10–30 Gb (up to 150 Gb for 5)
PromethION 24	Up to 24	~12000	~3000	100–200 Gb (up to 290 Gb) (up to 7 Tb for 24)
PromethION 2 Integrated (P2i)	Up to 2	~12000	~3000	Up to 290 Gb

Gridion Plateforme OMICS MIO

Disposition d'une Grille « Flow cell » & Contrôle MUX

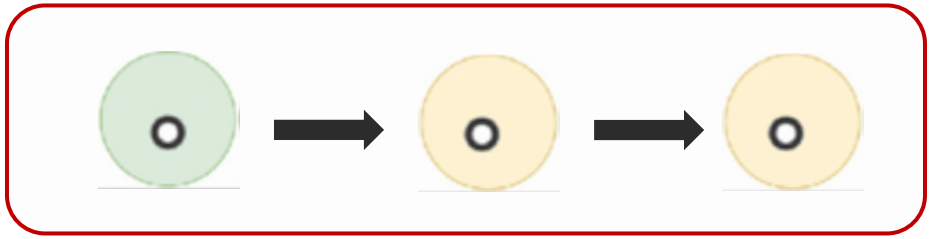
(MinIon/Gridion)



Carlos de Lannoy *et al.*, 2017

512 canaux

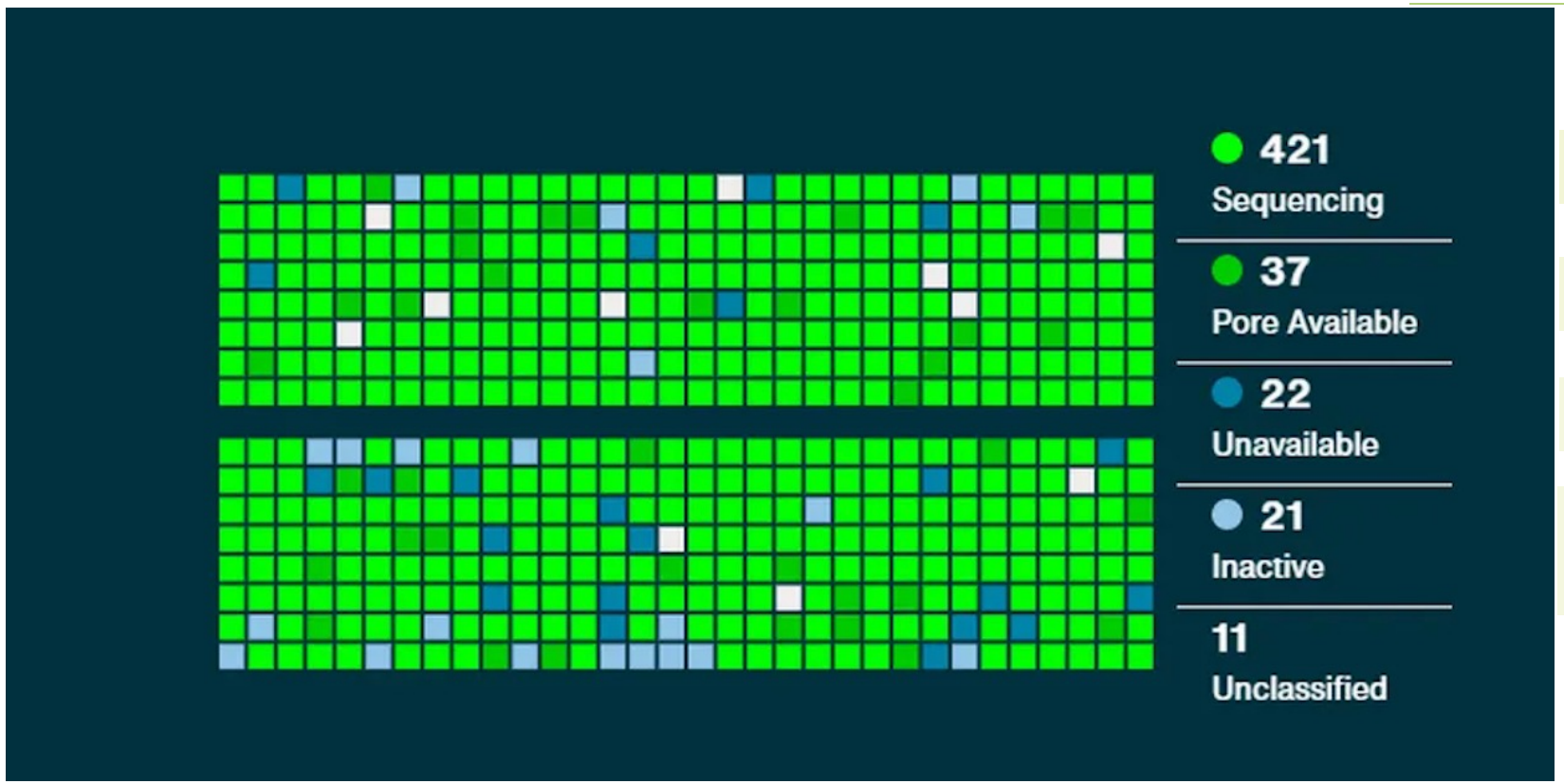
- 4 puits/canal
- 1 pore/puits (10 μm)



MUX contrôle

→ **Insertion incorrecte du pore**  **défaut de membrane**  **ou pores bloqués** 

Carte des nanopores



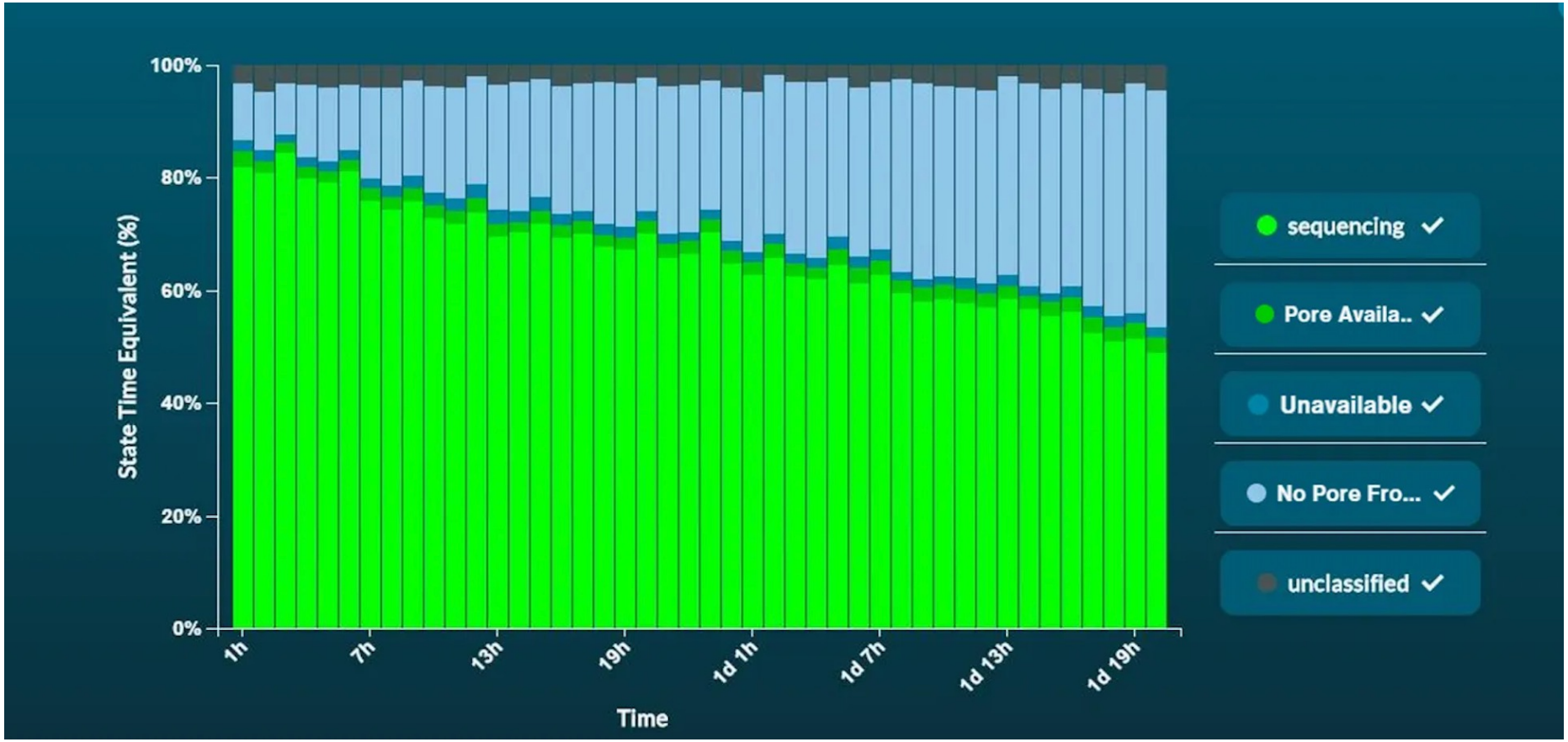
Actif séquençage

Reserve active mobilisable

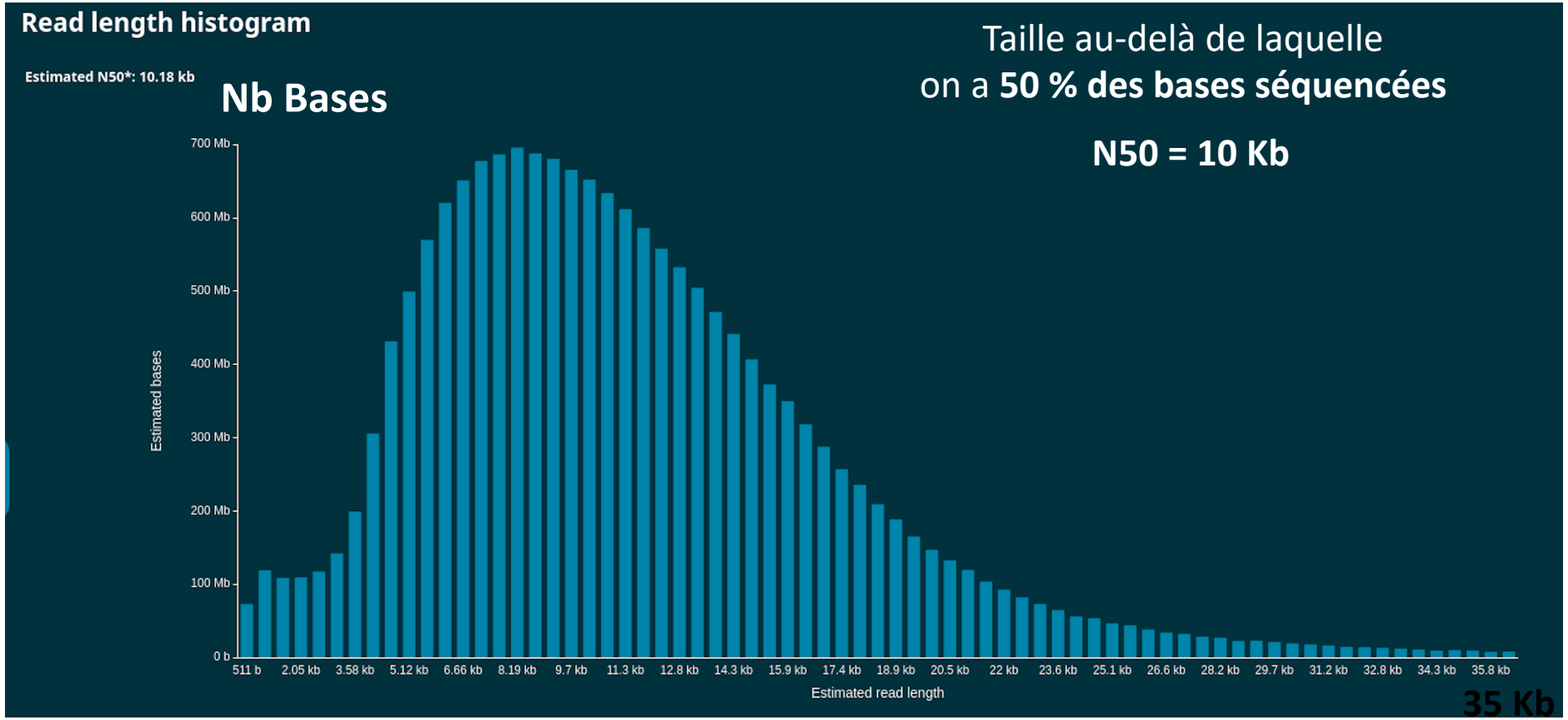
Non utilisables (Irrécupérables)

Pore en “veille” (Potentiellement récupérable)

Activité des Pores en fonction du Temps de Séquençage

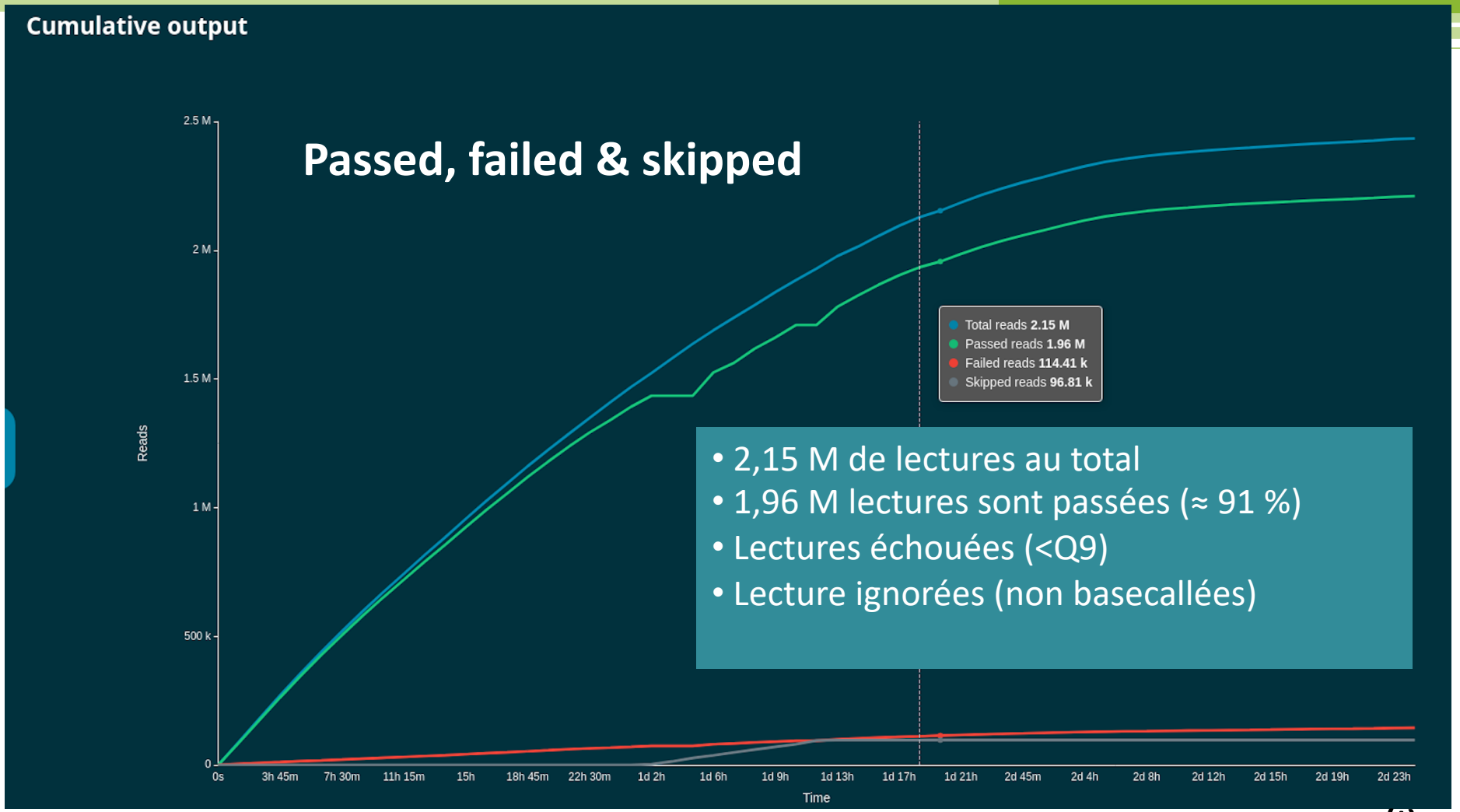


Distribution de la Longueur des Lectures

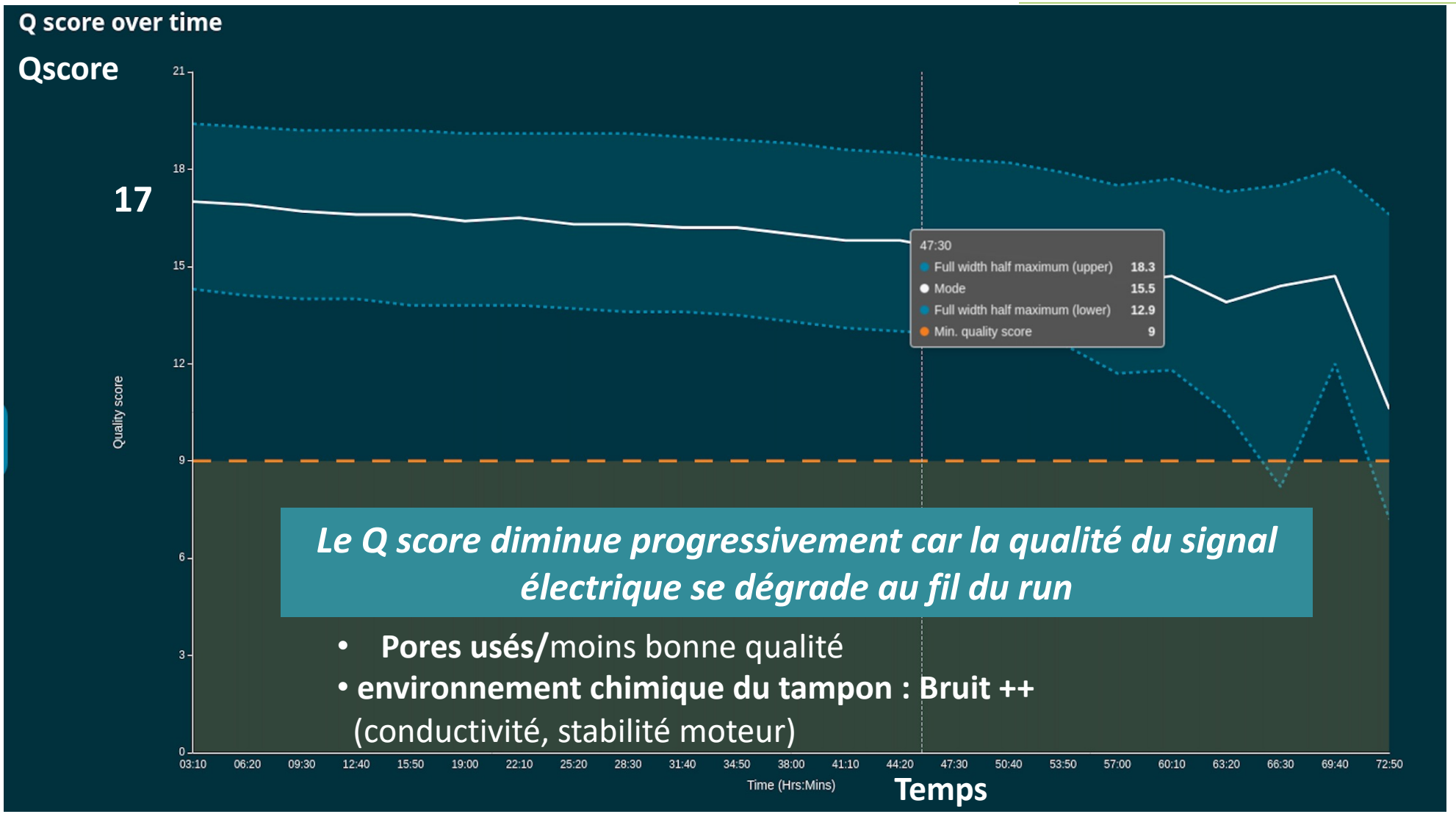


Longueur des lectures

Qualité des Lectures & Tri



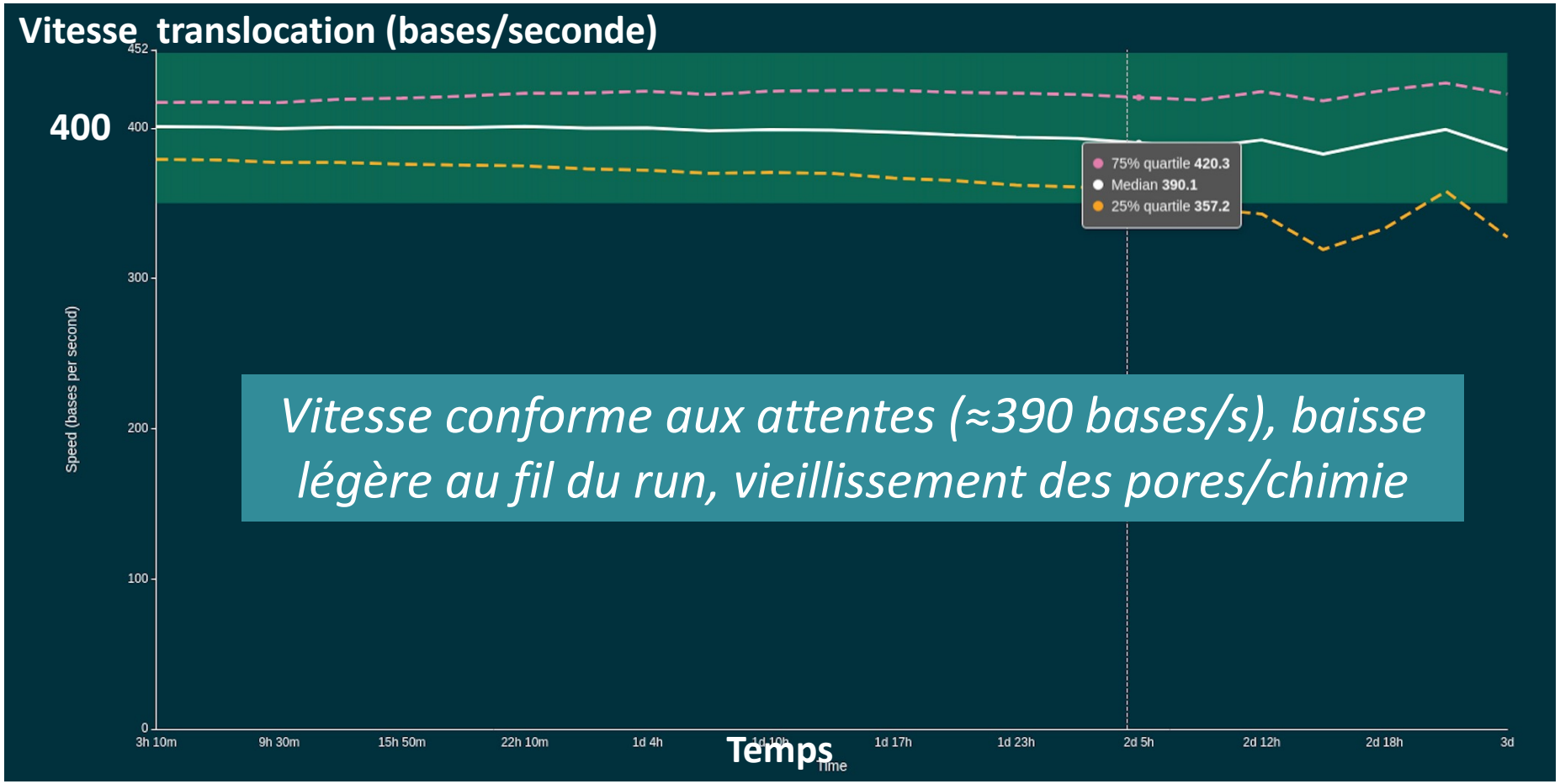
Variation du Qscore en Fonction du Temps de Run



Le Q score diminue progressivement car la qualité du signal électrique se dégrade au fil du run

- **Pores usés/**moins bonne qualité
- **environnement chimique du tampon : Bruit ++**
  (conductivité, stabilité moteur)

Vitesse de Passage des Bases dans les Nanopores



Vitesse conforme aux attentes (≈390 bases/s), baisse légère au fil du run, vieillissement des pores/chimie

Approche Analytique 16S Metabarcoding Nanopore

- **Preprocessing Nanopre différent d'Illumina**!!

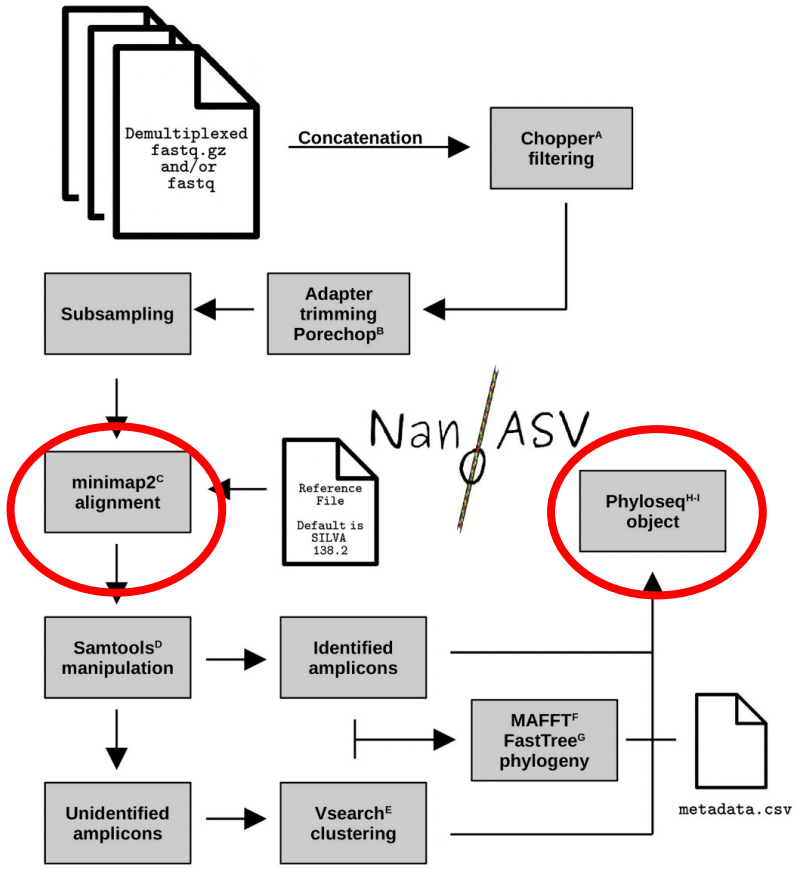
→ Pas d'assemblage (séquences pleine longueur, 1500 pb)

→ Pas le même taux d'erreur → ASV? OTU? ??

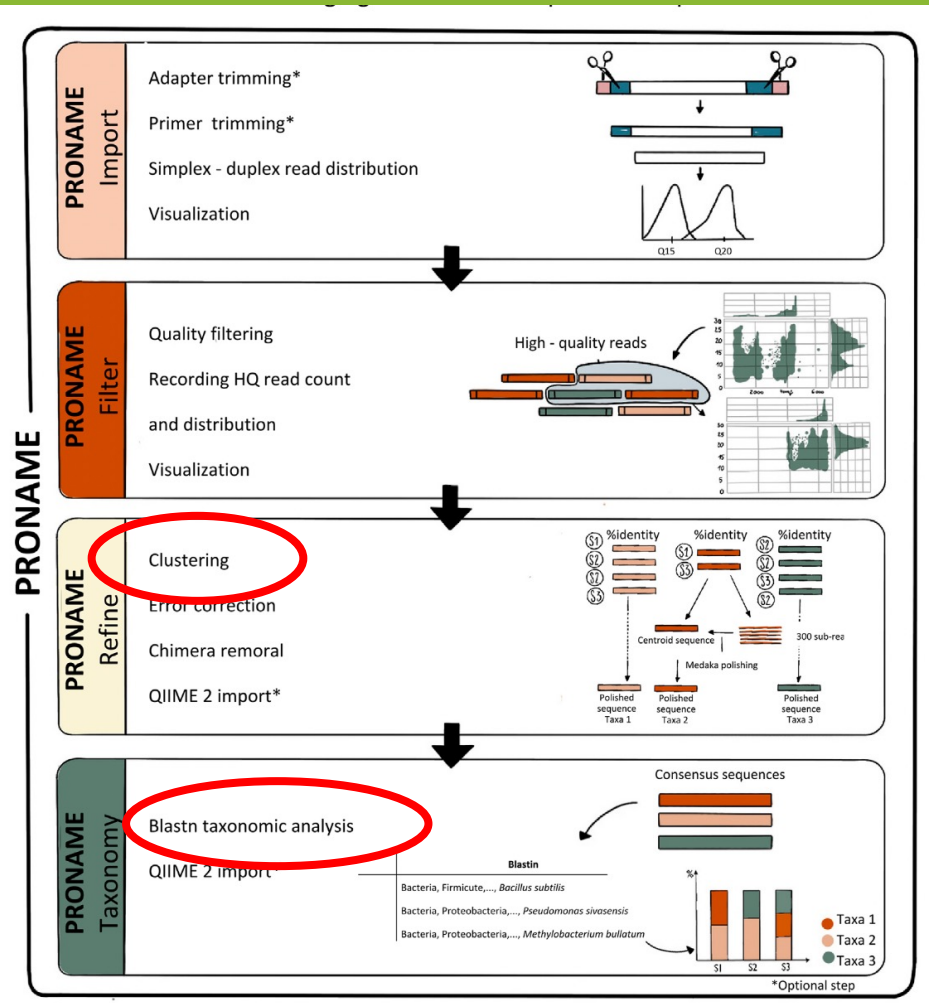
- Pas de « **Gold standard** » au niveau de l'approche comme Illumina

- **EPI2ME (soft Nanopore) → NO WAY!**

16S Nanopore Pipelines Récents



NanoASV - Cousson et al., 2025



PRONAME- Dubois et al., 2024

No/Partial Clustering vs. Full Clustering

Preprocessing Nanopore - EMU – Phyloseq

« Real » Sample Data

Mock Community = Etalon Interne

= Aide à la décision pour Quality trimming

NANOSTATS

Primer trimming
Cutadapt

Adaptator trimming
PorechopABI

Quality trimming
Seqkit

Chimera search*
Vsearch

Taxonomy Classification EMU
KD. Curry et al., Nat. 2022

Phyloseq Object
α, β Diversité

MultiQC processing output

Preprocessing Nanopore – Primer trimming

« Real » Sample Data

Mock Community = Etalon Interne

NANOSTATS

Primer trimming

Cutadapt

Tous **les reads commencent par la MÊME séquence** (le primer)

→ les reads semblent artificiellement plus proches

→ Leur absence peut également indiquer des reads de mauvaise qualité ou tronqués

Preprocessing Nanopore – Adaptator trimming

« Real » Sample Data

Mock Community = Etalon Interne

NANOSTATS

Primer trimming
Cutadapt

Adaptator trimming
PorechopABI

Adapter trimming – Nanopore (Porechop_ABI)

L'adaptateur (séquences artificielles) permet la liaison au système Nanopore

- Sert **d'interface entre l'ADN et la protéine** motrice,
- Rôle dans la **reconnaissance** du fragment et **augmente la capture (ADN)** par le pore

→ **Le trimming permet d'éliminer les artefacts de séquençage et d'éviter des erreurs d'assignation en aval**

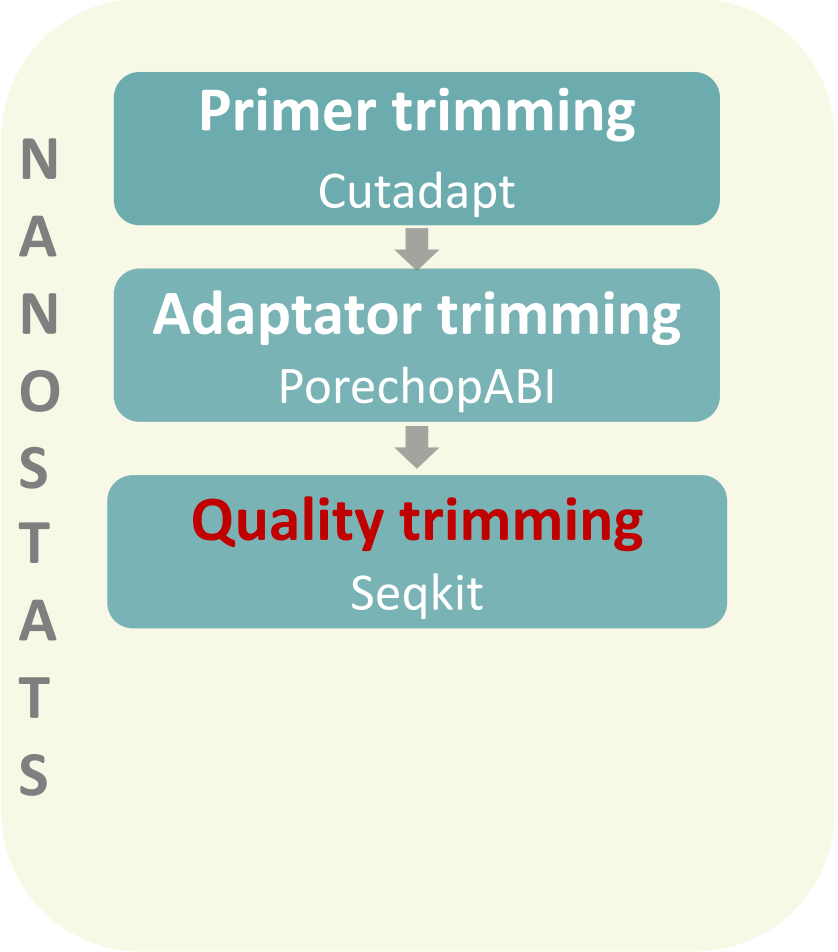
Cas : **ADN1 + adaptateur + ADN2**!
Adaptateur en milieu de séquence = chimères

Trimming :
✓ liste d'adaptateurs connus (séquences ONT)
✓ recherche dans les reads

Preprocessing Nanopore – Quality Trimming

« Real » Sample Data

Mock Community = Etalon Interne



Filtrer les reads en fonction de **leur score de qualité** (Q-score), définir longueur minimale et maximale

→ Garder uniquement les reads dont la **qualité moyenne ≥ seuil**

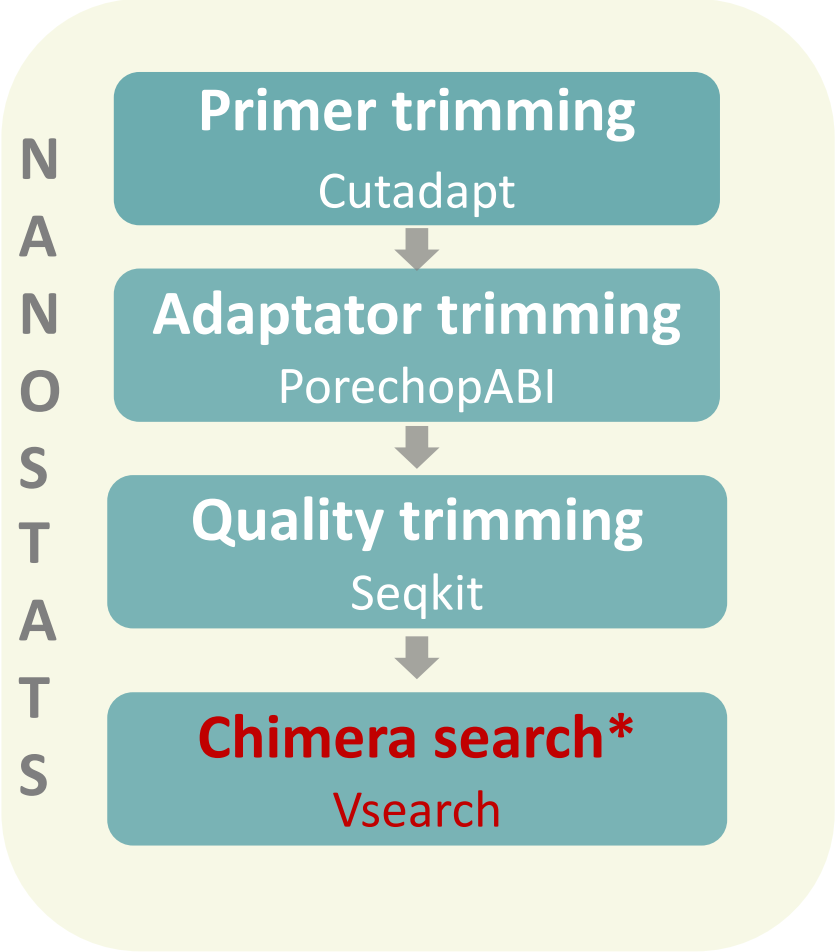
→ minimum ≤ **Longueur** ≤ maximum

Comment choisir ce seuil ??? → Mock

Preprocessing Nanopore - Chimères

« Real » Sample Data

Mock Community = Etalon Interne



Une chimère = une séquence artificielle formée d'au moins deux ADN différents « collés » ensemble

- Amplification incomplète
- Ré-hybridation entre fragments

→ détection de séquences artificielles issues de la fusion de fragments différents (base données)

Preprocessing Nanopore : Output (multiQC)

Sample Name	Median length	Read N50	Median Qual	# Reads (K)	Total Bases (Mb)
barcode13_1_input	1 582 bp	1 584 bp	16.4	328.4 K	510.6 Mb
barcode13_2_noPrimer	1 503 bp	1 504 bp	17.5	326.1 K	482.2 Mb
barcode13_3_noAdaptator	1 452 bp	1 453 bp	20.7	324.8 K	462.1 Mb
barcode13_4_Q17_qual	1 454 bp	1 455 bp	21.6	250.6 K	363.3 Mb
barcode13_4_Q22_qual	1 454 bp	1 455 bp	24.0	112.9 K	163.6 Mb

Q17 ou Q22?
Comment choisir le seuil qualité à appliquer ?
Meilleure Classification à Q22 ???

Preprocessing Nanopore - EMU – Phyloseq

« Real » Sample Data

Mock Community = Etalon Interne

NANOSTATS

Primer trimming
Cutadapt

Adaptator trimming
PorechopABI

Quality trimming
Seqkit

Chimera search*
Vsearch

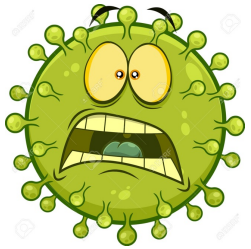
Taxonomy Classification EMU
KD. Curry et al., Nat. 2022

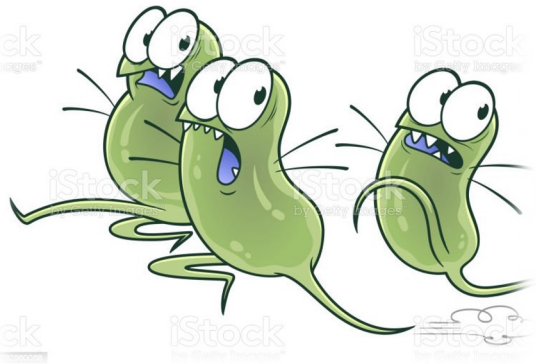
Phyloseq Object
α, β Diversité

EMU Classification : Expectation–Maximization for Ultra-long reads

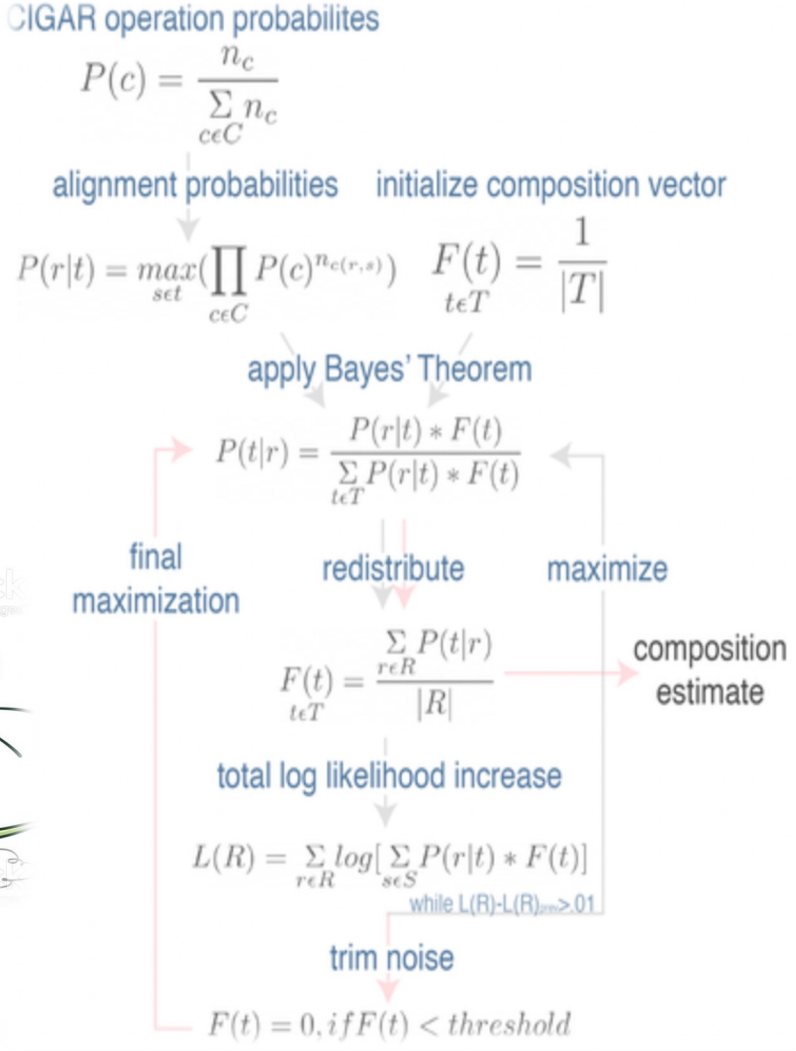
KD. Curry *et al.*, Nat. Method, 2022

Postulat : Une lecture peut correspondre à **plusieurs espèces proches!**
→ l'assigner directement peut donner des erreurs!!





https://github.com/treangenlab/emu



Tenir compte de l'incertitude d'assignation des reads

EMU Classification : Expectation–Maximization for Ultra-long reads

KD. Curry *et al.*, Nat. Method, 2022

Idée Générale :

- EMU **ne "classe" pas chaque read** (i.e. **1 read -> un taxon**) → il estime la composition globale du microbiome

Répond à la question :

- Quelles espèces sont présentes, et en quelle proportion ?

Pourquoi ?

- reads longs
- erreurs (indels)
- un read peut correspondre à plusieurs espèces

→ impossible d'assigner chaque read de façon **fiable à UNE espèce**

EMU Classification : Expectation–Maximization for Ultra-long reads

KD. Curry *et al.*, Nat. Method, 2022

On ne dit pas : ce read = cette espèce

MAIS : ce read pourrait venir de plusieurs espèces

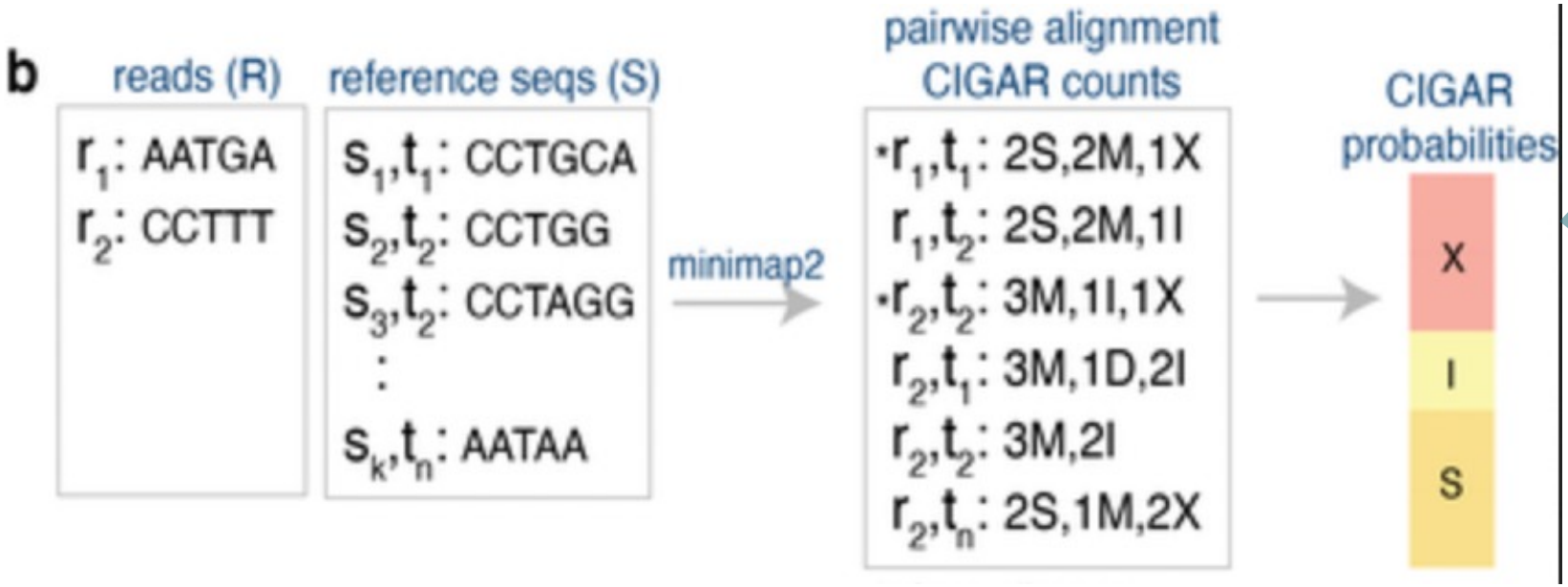
read1 → espèces possibles read1

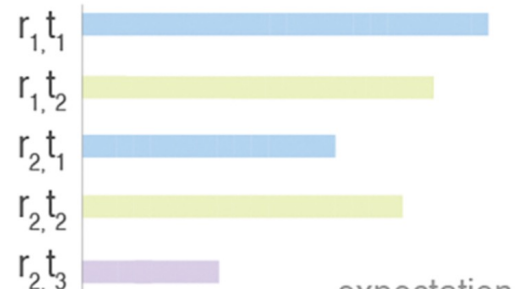
→ E. coli (60%), Shigella (30%), C. koseri (10%)

et on ajuste progressivement

1- Alignement Initial & Probabilité

Sélection basée sur un seuil de qualité d'alignement (CIGAR, % id, …)





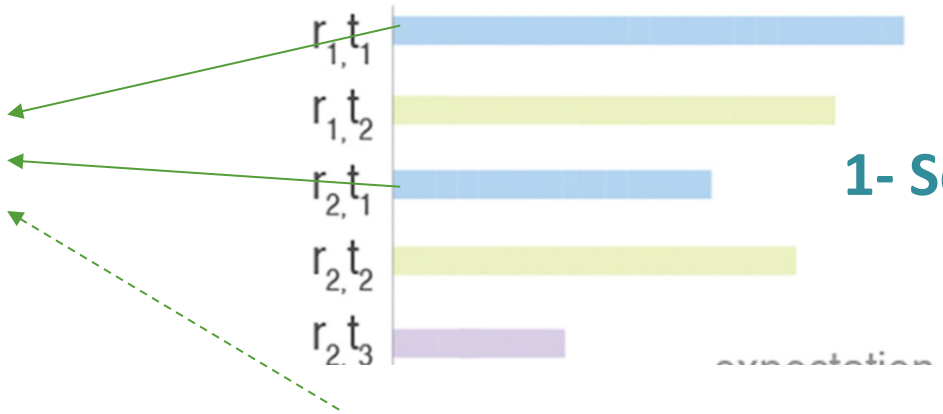
EMU : A chaque lecture une **liste de taxons candidats!**
Et une probabilité d'origine (t1,t2…)

→ **PAS DE BEST MATCH!!**

probabilité d'origine

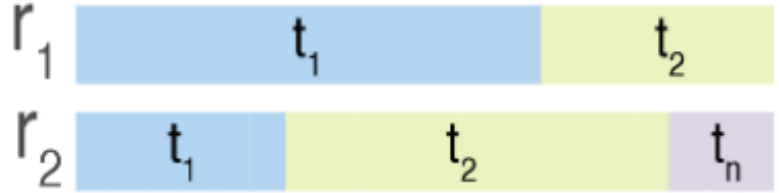
2- Expectation Step : Contributions Pondérées par Lecture

Abondance Globale t1



1- Score Alignement → Probabilité d'origine

2- Expectation Step : Fractionnement en contributions partielles

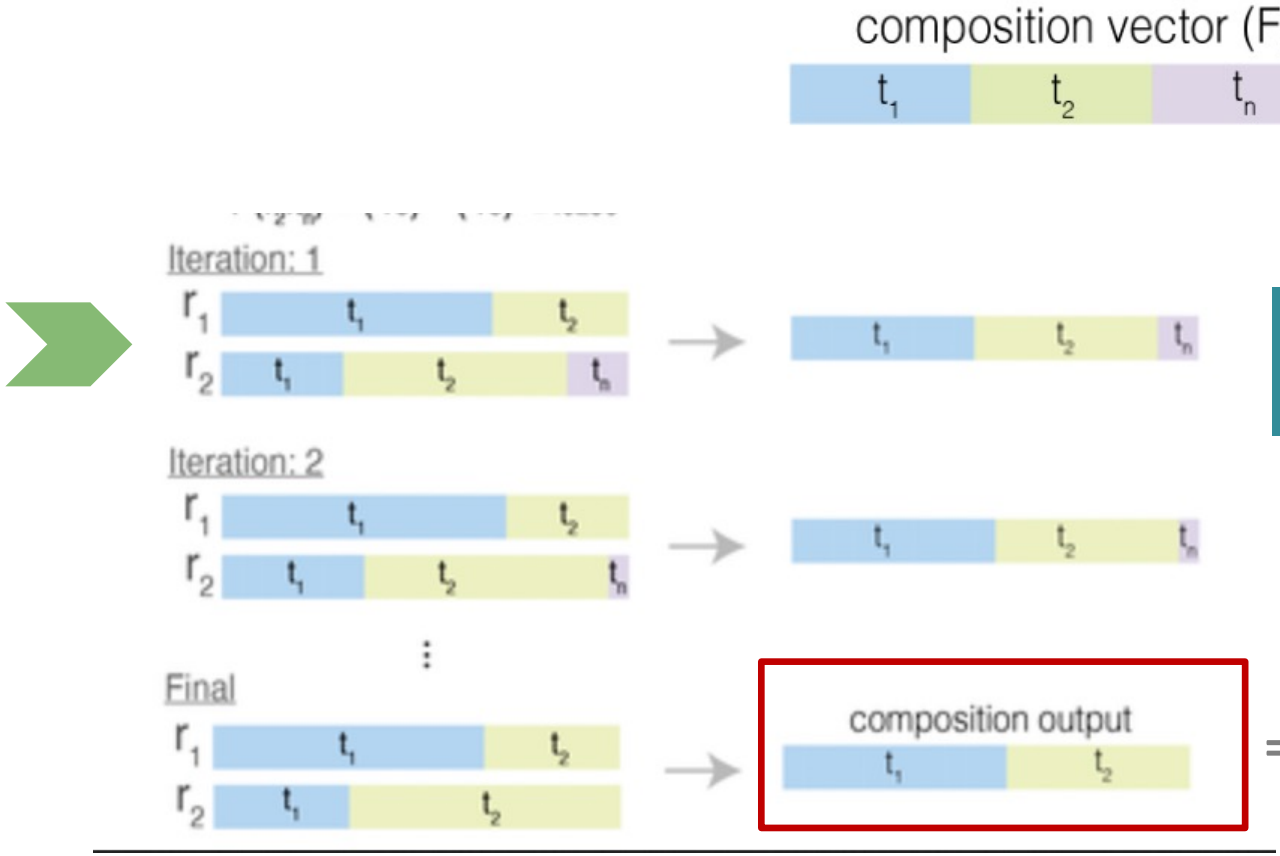


Pondération EMU : E-step

Une lecture ambiguë est distribuée sous forme de **fractions de probabilité ($P_{origine}$, abondance globale, normalisation)**

Maximisation-Step : EMU : Ajustement Abondance Globale

Abondance Globale initiale égale pour tous les taxons



Ajustement Abondance globale à chaque itération

= Taxonomie + Abondance relative

M-step : Sommer toutes les contributions pour recalculer les **abondances globales** des taxons

Conclusion EMU

✓ **Robuste au bruit Nanopore**
✓ Gère erreurs Nanopore
✓ Gère ambiguïtés taxonomiques

EMU plus robuste que “assignation brute”

Estimer les abondances des espèces en tenant compte de **l’incertitude d’assignation des reads**, ce qui permet une estimation plus robuste de la composition microbienne.

Preprocessing Nanopore - EMU – Phyloseq

« Real » Sample Data

Mock Community = Etalon Interne

???

NANOSTATS

Primer trimming
Cutadapt

Adaptator trimming
PorechopABI

Quality trimming
Seqkit

Chimera search*
Vsearch

Taxonomy Classification EMU
KD. Curry et al., Nat. 2022

Le +++ du script

Phyloseq Object
α, β Diversité

MultiQC processing output

Rappel

Sample Name	Median length	Read N50	Median Qual	# Reads (K)	Total Bases (Mb)
barcode13_1_input	1 582 bp	1 584 bp	16.4	328.4 K	510.6 Mb
barcode13_2_noPrimer	1 503 bp	1 504 bp	17.5	326.1 K	482.2 Mb
barcode13_3_noAdaptator	1 452 bp	1 453 bp	20.7	324.8 K	462.1 Mb
barcode13_4_Q17_qual	1 454 bp	1 455 bp	21.6	250.6 K	363.3 Mb
barcode13_4_Q22_qual	1 454 bp	1 455 bp	24.0	112.9 K	163.6 Mb

Q17 ou Q22?
Comment choisir le seuil qualité à appliquer ?
Meilleure Classification à Q22 ???

Mock community = Choisir le Qscore à appliquer…

Zymobiotic Standard

Species	Theoretical Composition (%)				
	Genomic DNA	16S Only[1]	16S & 18S[1]	Genome Copy[2]	Cell Number[3]
Pseudomonas aeruginosa	12	4.2	3.6	6.1	6.1
Escherichia coli	12	10.1	8.9	8.5	8.5
Salmonella enterica	12	10.4	9.1	8.7	8.8
Lactobacillus fermentum	12	18.4	16.1	21.6	21.9
Enterococcus faecalis	12	9.9	8.7	14.6	14.6
Staphylococcus aureus	12	15.5	13.6	15.2	15.3
Listeria monocytogenes	12	14.1	12.4	13.9	13.9
Bacillus subtilis	12	17.4	15.3	10.3	10.3
Saccharomyces cerevisiae	2	NA	9.3	0.57	0.29
Cryptococcus neoformans	2	NA	3.3	0.37	0.18

Zymobiotic Standard II (log Distribution)

Species	Theoretical Composition (%)				
	Genomic DNA	16S Only[1]	16S & 18S[1]	Genome Copy[2]	Cell Number[3]
Listeria monocytogenes	89.1	95.9	91.9	94.8	94.9
Pseudomonas aeruginosa	8.9	2.8	2.7	4.2	4.2
Bacillus subtilis	0.89	1.2	1.1	0.7	0.7
Saccharomyces cerevisiae	0.89	NA	4.1	0.23	0.12
Escherichia coli	0.089	0.069	0.066	0.058	0.058
Salmonella enterica	0.089	0.07	0.067	0.059	0.059
Lactobacillus fermentum	0.0089	0.012	0.012	0.015	0.015
Enterococcus faecalis	0.00089	0.00067	0.00064	0.001	0.001
Cryptococcus neoformans	0.00089	NA	0.0014	0.00015	0.00007
Staphylococcus aureus	0.000089	0.0001	0.0001	0.0001	0.0001

Comparer valeurs Théoriques vs. Expérimentales (post séquençage – Qscore variation)
Choisir Qscore qui minimise les différences!!

Utilisation de métriques pour guider le choix

Le RMSE (Root Mean Square Error) : Ecart moyen entre les **valeurs prédites et réelles** (=erreur) en donnant plus de poids aux grandes erreurs grâce à la **mise au carré des écarts**

- Plus le RMSE est petit, plus les prédictions sont proches de la réalité.

La MAE (Mean Absolute Error) mesure l'écart moyen entre les valeurs prédites et les valeurs réelles, en prenant la **valeur absolue des erreurs**

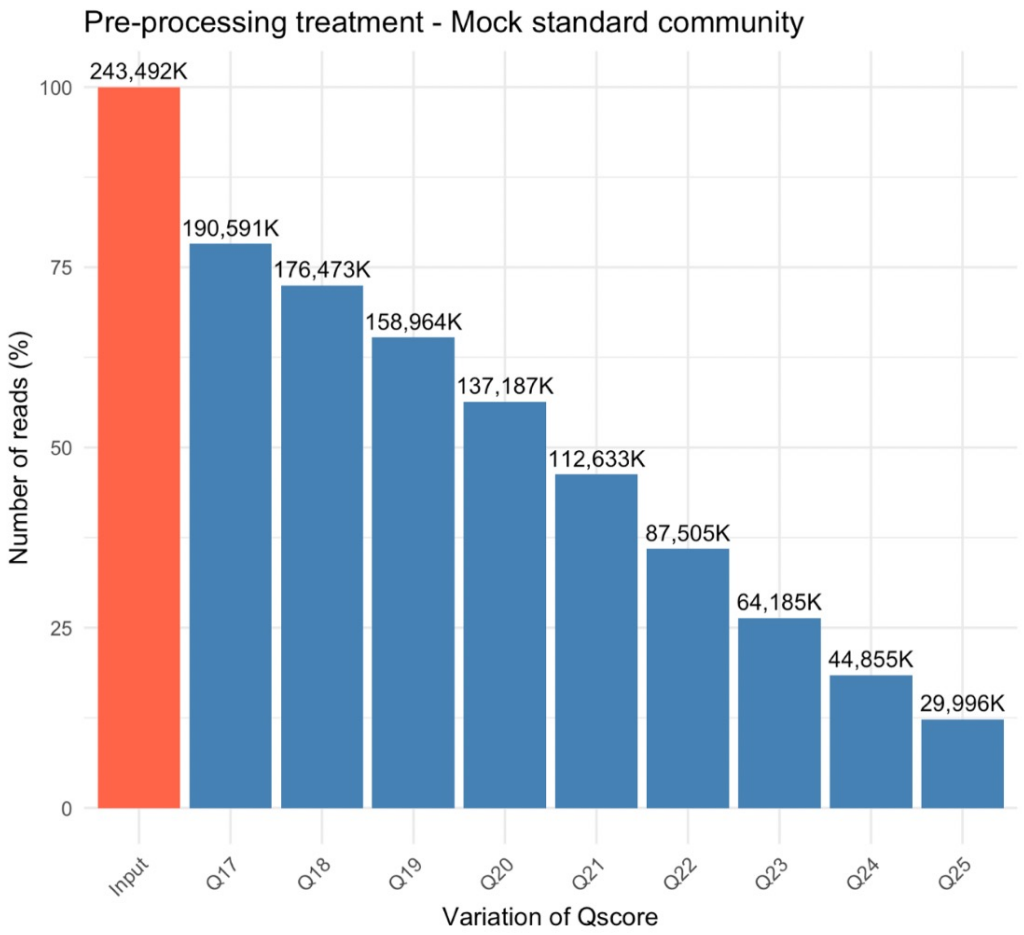
	MAE	**RMSE**
Calcul	valeur **absolue**	**carré** + racine
Sensibilité	erreurs traitées pareil	**grosses erreurs amplifiées**
Interprétation	plus intuitive	plus sensible aux outliers

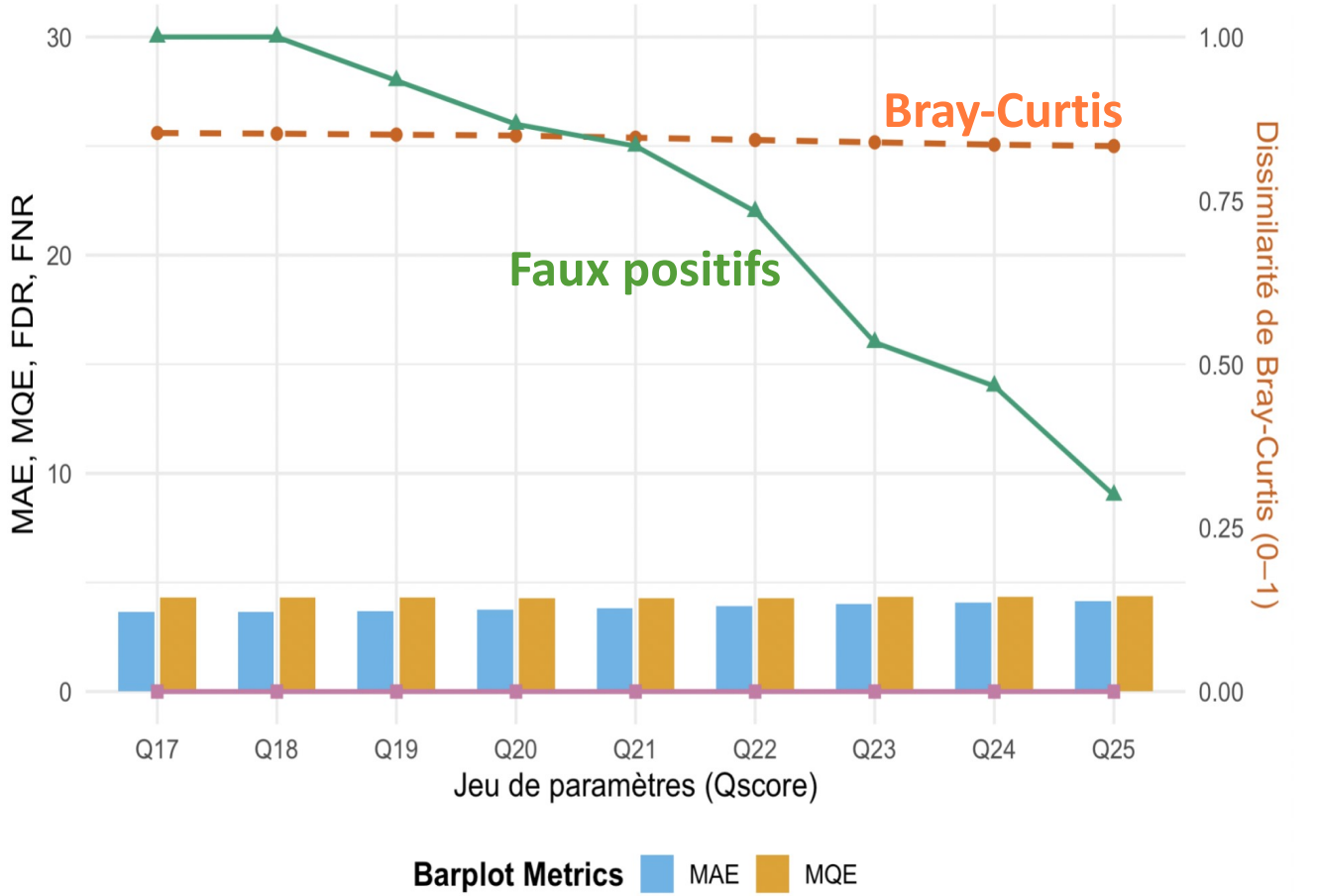
Bray-Curtis Similarité (1-Bray-Curtis)

La distance de Bray-Curtis permet de mesurer la dissimilarité entre deux profils (attendu vs observé) en tenant **compte des abondances relatives des espèces**.

Une valeur proche de 1 indiquant des profils similaires et proche de 0 des profils différents.

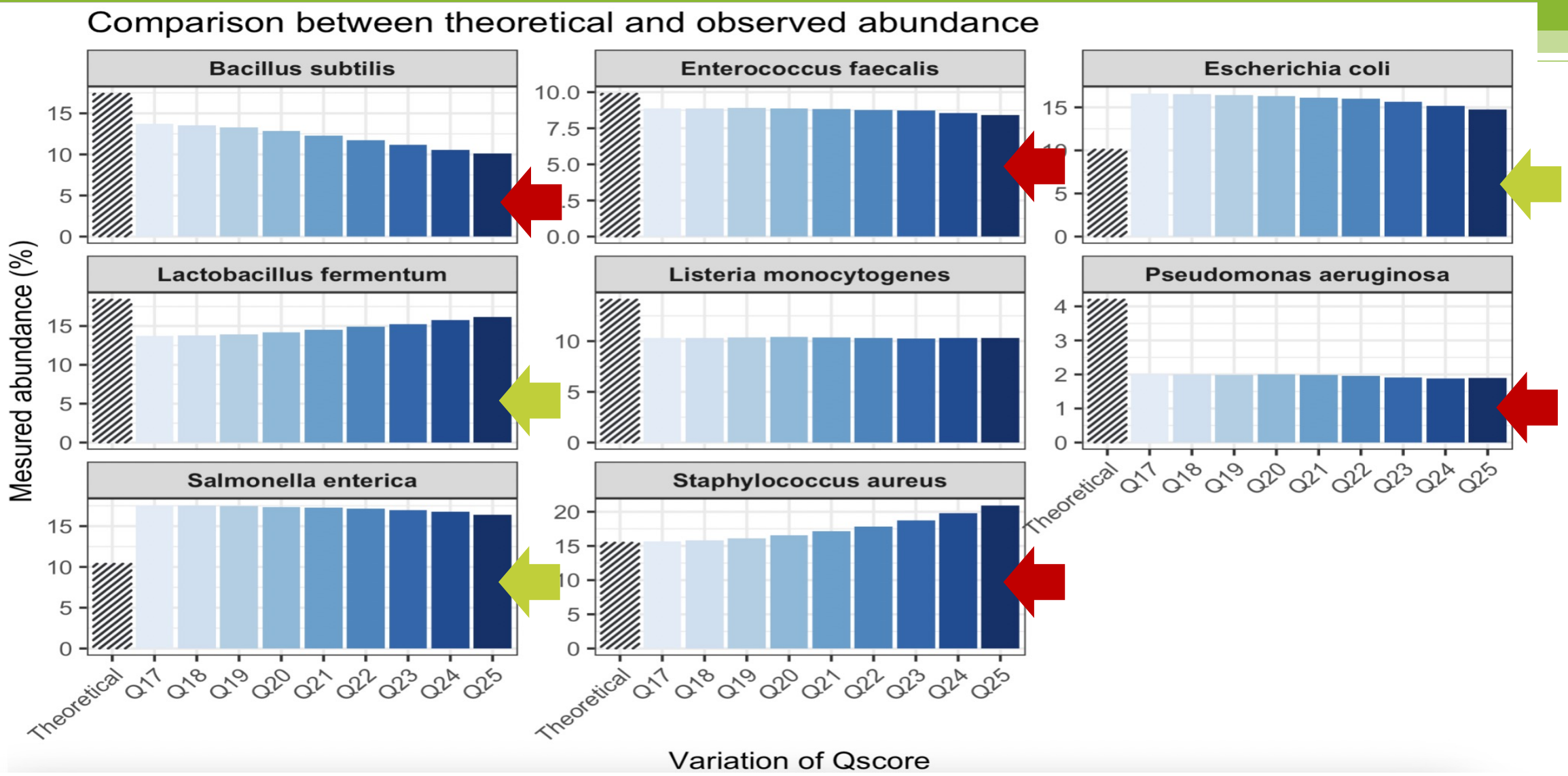
Variation Qscore : Adéquation Théorie vs. Expérimental



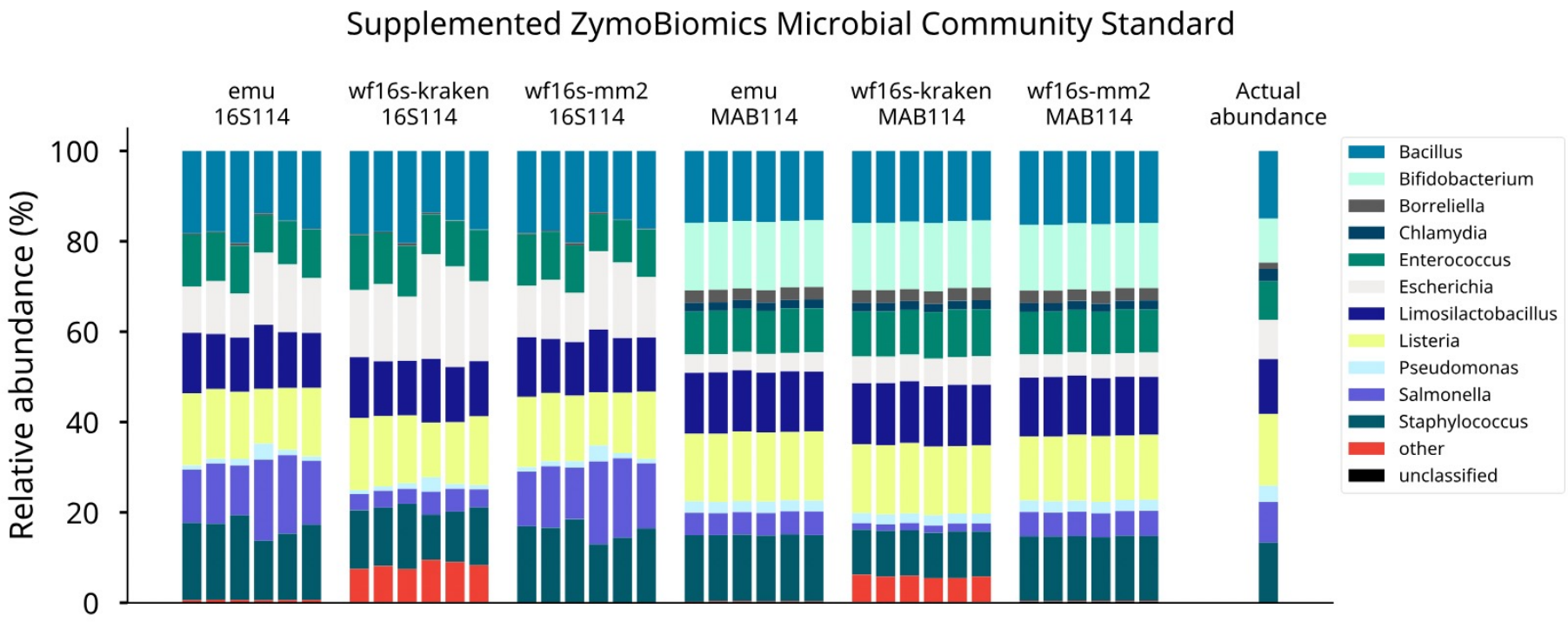


Erreur absolue moyenne = MAE
Racine de l'erreur quadratique moyenne = RMSE (= MQE)
Test χ^2 d'adéquation; G-test

Abondance relative & Qscore

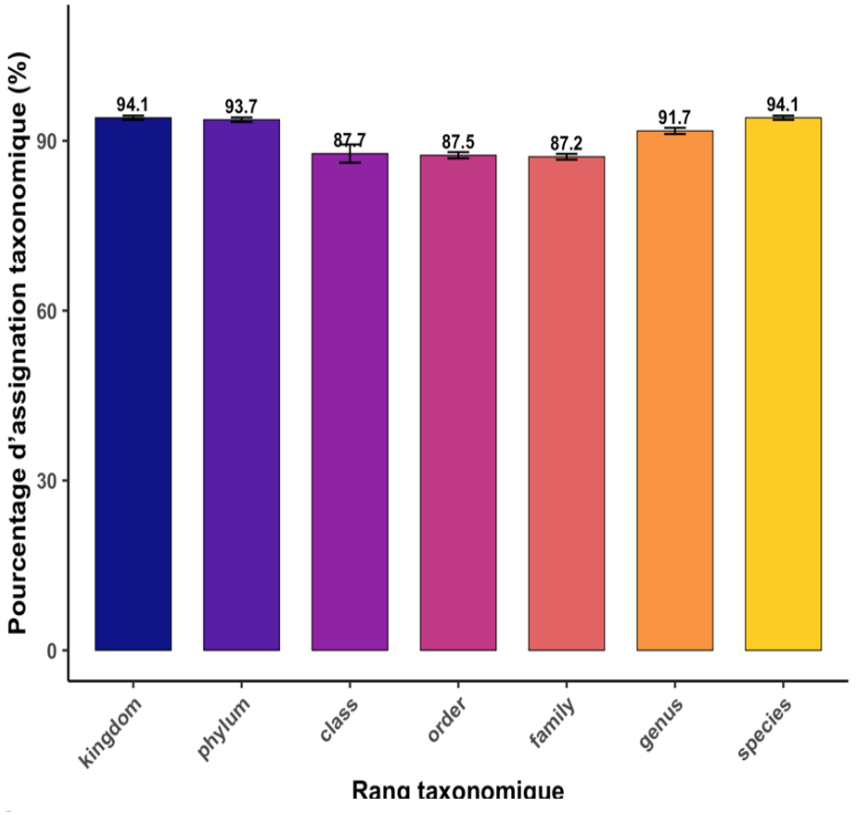


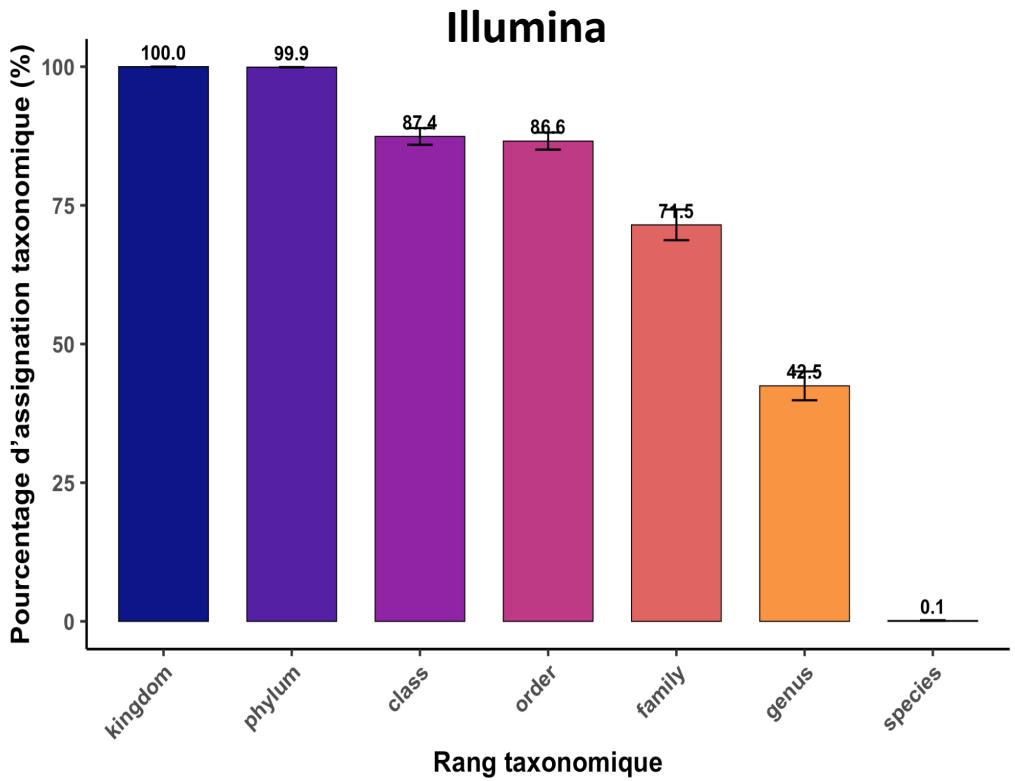
Derniers Tests Nanopore sur Mock (sept 2025)



→ Minimap2/Emu meilleurs pour classification

Impact sur la Classification au Rang de l'Espèce





Metabarcoding via Nanopore

- Séquençage complet du marqueur
- Taux erreur moyen < 1% (> Q20)
- EMU classification
- Classification au rang Espèce ++
- ~ 20 Gb/Flowcell → 13 M lectures/sample
- → 0,27 M lectures/ 50 samples

SUPPORT POUR TP

MetaDonnées du TP = MAPFILE

- Samples : Eaux de la mer Méditerranée (proche Marseille)
- Collectée au niveau de deux **Stations** : A et B (= Variables catégories)
- Différentes **profondeurs** : 75m, 100m, 200m, 300m et 500m (= Variables catégories)
- **Paramètres environnementaux** (O2, NO3, pH…) expliquant la structure des communautés microbiennes (= variables continues) : TOTALEMENT INVENTES!!!!

- O2, NO3, NH4 : mg/L
- Température : °C
- Conductivité : µS/cm
- pH : sans unité

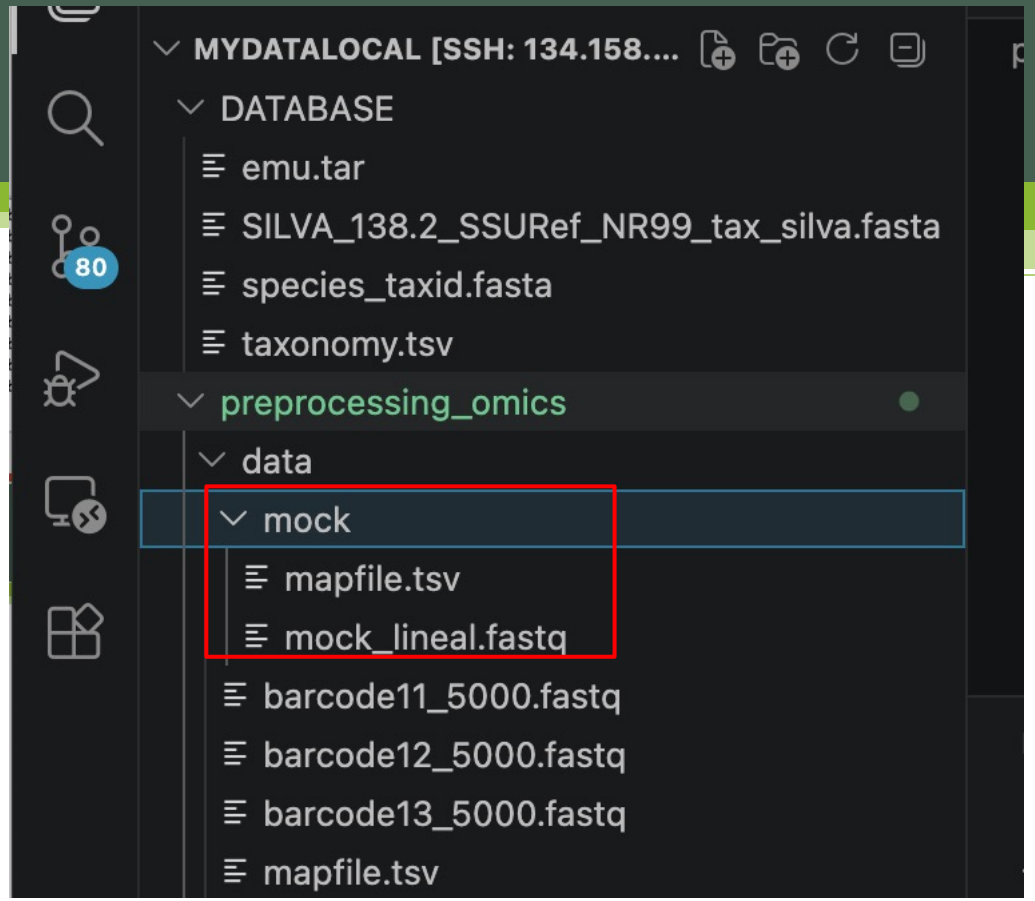
Ca va ressembler à cela & c'est dans physeq@sam_data

```
> env
                                SampleID Technology KIT Depth Station      Name Depths Depth_cat  O2
barcode13_combine_Q22 barcode13_combine_Q22   Nanopore   A  300m     ST_A ST_A_300m    300       mid 128
barcode14_combine_Q22 barcode14_combine_Q22   Nanopore   A  500m     ST_A ST_A_500m    500      deep  82
barcode17_combine_Q22 barcode17_combine_Q22   Nanopore   B   75m     ST_B  ST_B_75m     75   shallow 232
barcode18_combine_Q22 barcode18_combine_Q22   Nanopore   B  100m     ST_B ST_B_100m    100   shallow 221
barcode19_combine_Q22 barcode19_combine_Q22   Nanopore   B  200m     ST_B ST_B_200m    200       mid 171
barcode20_combine_Q22 barcode20_combine_Q22   Nanopore   B  300m     ST_B ST_B_300m    300       mid 129
barcode21_combine_Q22 barcode21_combine_Q22   Nanopore   B  500m     ST_B ST_B_500m    500      deep  82
barcode22_combine_Q22 barcode22_combine_Q22   Nanopore   A   75m     ST_A  ST_A_75m     75   shallow 228
barcode23_combine_Q22 barcode23_combine_Q22   Nanopore   A  100m     ST_A ST_A_100m    100   shallow 208
barcode24_combine_Q22 barcode24_combine_Q22   Nanopore   A  200m     ST_A ST_A_200m    200       mid 166
                       NO3  NH4  NO2 Salinity   pH Turbidity Pollutant
barcode13_combine_Q22 18.5 0.94 0.11    38.32 8.06       1.8      9.44
barcode14_combine_Q22 27.6 0.57 0.05    37.88 7.98       2.1      9.33
barcode17_combine_Q22  2.4 0.44 0.04    38.41 8.12       1.5      9.30
barcode18_combine_Q22  4.2 0.41 0.13    38.05 8.03       2.3      9.30
barcode19_combine_Q22 10.8 0.52 0.17    37.72 7.95       1.9      9.00
barcode20_combine_Q22 18.1 0.86 0.09    38.28 8.10       2.0      8.84
barcode21_combine_Q22 28.5 0.43 0.03    37.95 8.01       1.7      8.57
barcode22_combine_Q22  2.0 0.40 0.06    38.10 7.92       2.2      9.74
barcode23_combine_Q22  4.8 0.69 0.11    37.83 8.08       1.6      9.83
barcode24_combine_Q22  9.9 0.98 0.19    38.36 8.00       1.8      9.66
```

Données & Mock

Dans dossier mock

- Mock_lineal.fastq
- Mapfile.tsv



Toujours mettre données séquençage (fastq) de votre mock et mapfile correspondant dans ce dossier « mock »

Attention : Mock community → lineal or log

Zymobiotic Standard

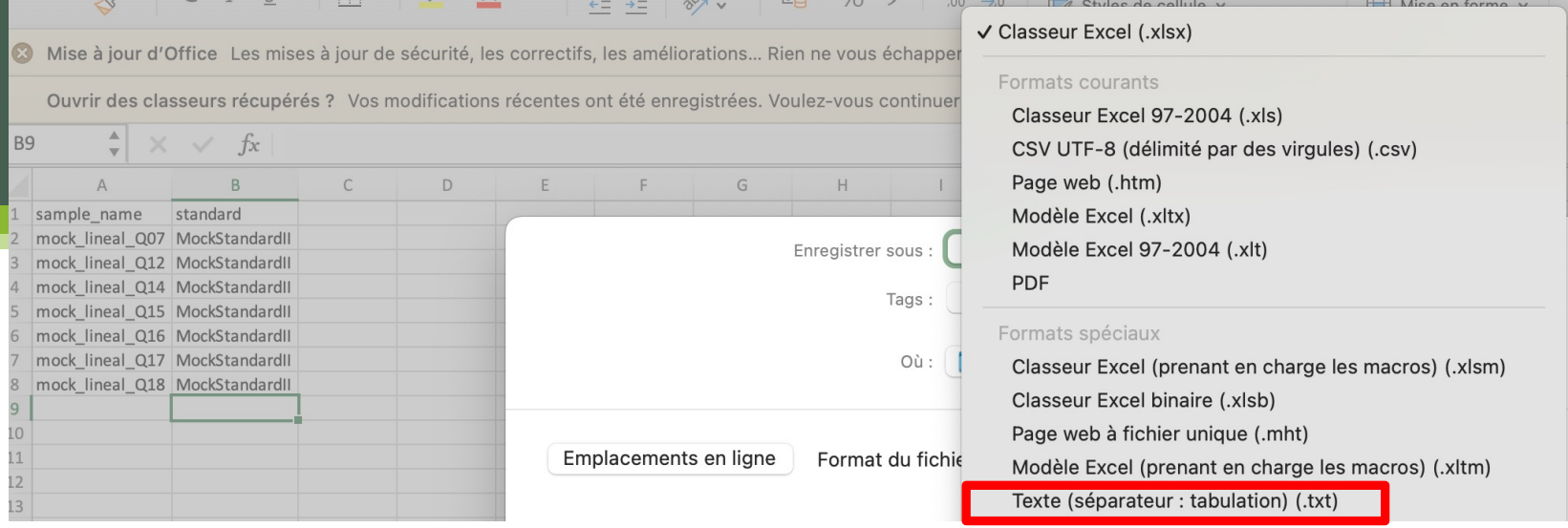
Species	Theoretical Composition (%)				
	Genomic DNA	16S Only[1]	16S & 18S[1]	Genome Copy[2]	Cell Number[3]
Pseudomonas aeruginosa	12	4.2	3.6	6.1	6.1
Escherichia coli	12	10.1	8.9	8.5	8.5
Salmonella enterica	12	10.4	9.1	8.7	8.8
Lactobacillus fermentum	12	18.4	16.1	21.6	21.9
Enterococcus faecalis	12	9.9	8.7	14.6	14.6
Staphylococcus aureus	12	15.5	13.6	15.2	15.3
Listeria monocytogenes	12	14.1	12.4	13.9	13.9
Bacillus subtilis	12	17.4	15.3	10.3	10.3
Saccharomyces cerevisiae	2	NA	9.3	0.57	0.29
Cryptococcus neoformans	2	NA	3.3	0.37	0.18

Zymobiotic Standard II (log Distribution)

Species	Theoretical Composition (%)				
	Genomic DNA	16S Only[1]	16S & 18S[1]	Genome Copy[2]	Cell Number[3]
Listeria monocytogenes	89.1	95.9	91.9	94.8	94.9
Pseudomonas aeruginosa	8.9	2.8	2.7	4.2	4.2
Bacillus subtilis	0.89	1.2	1.1	0.7	0.7
Saccharomyces cerevisiae	0.89	NA	4.1	0.23	0.12
Escherichia coli	0.089	0.069	0.066	0.058	0.058
Salmonella enterica	0.089	0.07	0.067	0.059	0.059
Lactobacillus fermentum	0.0089	0.012	0.012	0.015	0.015
Enterococcus faecalis	0.00089	0.00067	0.00064	0.001	0.001
Cryptococcus neoformans	0.00089	NA	0.0014	0.00015	0.00007
Staphylococcus aureus	0.000089	0.0001	0.0001	0.0001	0.0001

Ne pas vous tromper!!!

& mapfile....



FORMAT OBLIGATOIRE Nomsample_Qscore

```
rocessing_omics/data/mock$ more mapfile.tsv
sample_name      standard
mock_lineal_Q07 MockStandardII
mock_lineal_Q12 MockStandardII
mock_lineal_Q14 MockStandardII
mock_lineal_Q15 MockStandardII
mock_lineal_Q16 MockStandardII
mock_lineal_Q17 MockStandardII
mock_lineal_Q18 MockStandardII
```

Correspondance avec -> QUAL_LIST_DEFAULT=(07 12 13 14 15 16 17 18) # liste de Qscores par défaut pour mock

Config.sh file for running mock community analysis

```
# --- Bases de données ---
EMU_DATABASE_DIR="/home/ubuntu/data/mydatalocal/DATABASE"
DATABASE_VSEARCH="$EMU_DATABASE_DIR/SILVA_138.2_SSURef_NR99_tax_silva.fasta"
# --- Options d'exécution ---
RUN_MOCK=true # activer le traitement « mock »

# --- Qualité et filtrage ---
QUAL=20 # score minimum Qscore
CPU=20 # nombre de threads
MINLEN=1400 # longueur minimale des reads
MAXLEN=1650 # longueur maximale des reads
QUAL_LIST_DEFAULT=(07 12 13 14 15 16 17) # liste de Qscores par défaut pour mock
QUAL_LIST=() # liste actuelle de Qscores ->  RIEN LA DEDANS
EMUtype="map-ont" # type EMU par défaut
# --- Test mock ---
MOCK_SCALE=  " lineal" # log, lineal, both (comment il trouve?? )
# --- Répertoires de sortie ---
OUTPUT_DIR="./RESULTS_mock"
```

Résultats mock : preprocessing des données : HTML

Dans le dossier RESULTS_mock/MULTIQC

NanoStat

Reports various statistics for long read dataset in FASTQ, BAM, or albacore sequencing summary format (supports NanoPack; NanoPlot, NanoComp). https://github.com/wdecoster/nanostat; https://github.com/wdecoster/nanoplot DOI: 10.1093/bioinformatics/bty149

Programs are part of the NanoPack family for summarising results of sequencing on Oxford Nanopore methods (MinION, PromethION etc.)

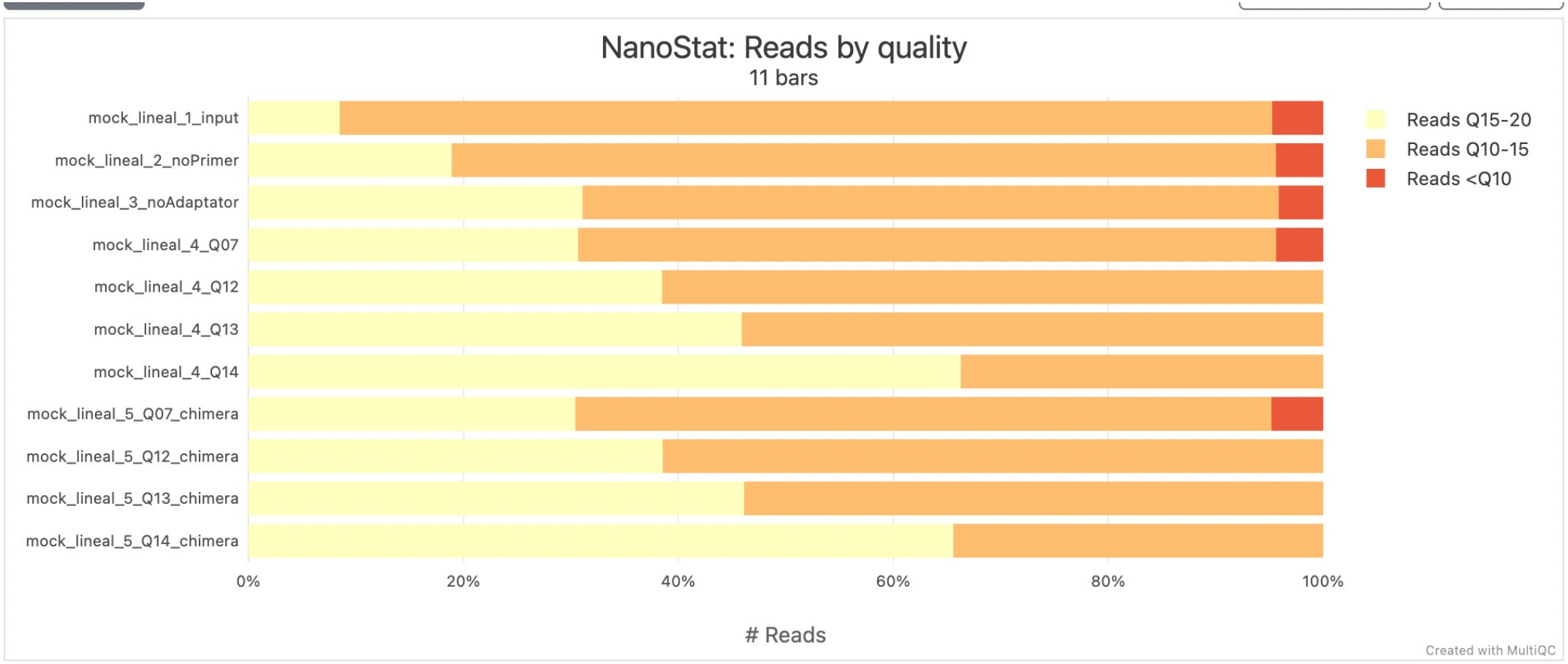
Summary Statistics (FASTQ)

Copy table | Configure columns | Scatter plot | Violin plot | Export as CSV... | Showing 11/11 rows and 5/7 columns. | Summarize table

Sample Name	Median length	Read N50	Median Qual	# Reads (K)	Total Bases (Mb)
mock_lineal_1_input	1604 bp	1605 bp	13.1	0.4 K	0.6 Mb
mock_lineal_2_noPrimer	1534 bp	1537 bp	13.5	0.4 K	0.6 Mb
mock_lineal_3_noAdaptator	1474 bp	1474 bp	13.8	0.4 K	0.6 Mb
mock_lineal_4_Q07	1475 bp	1475 bp	13.8	0.4 K	0.5 Mb
mock_lineal_4_Q12	1476 bp	1476 bp	14.4	0.3 K	0.4 Mb
mock_lineal_4_Q13	1478 bp	1478 bp	14.9	0.2 K	0.4 Mb
mock_lineal_4_Q14	1479 bp	1479 bp	15.2	0.2 K	0.3 Mb
mock_lineal_5_Q07_chimera	1475 bp	1475 bp	13.8	0.3 K	0.5 Mb
mock_lineal_5_Q12_chimera	1476 bp	1477 bp	14.4	0.3 K	0.4 Mb
mock_lineal_5_Q13_chimera	1478 bp	1478 bp	14.9	0.2 K	0.3 Mb
mock_lineal_5_Q14_chimera	1478 bp	1479 bp	15.2	0.2 K	0.2 Mb

NB : Le Read N50 correspond à la taille de lecture telle que 50 % des bases totales sont contenues dans des reads de cette taille ou plus longs.

Résultats mock : preprocessing des données : HTML



Résultats test mock : dans RESULTS_mock/METRICS/lineal

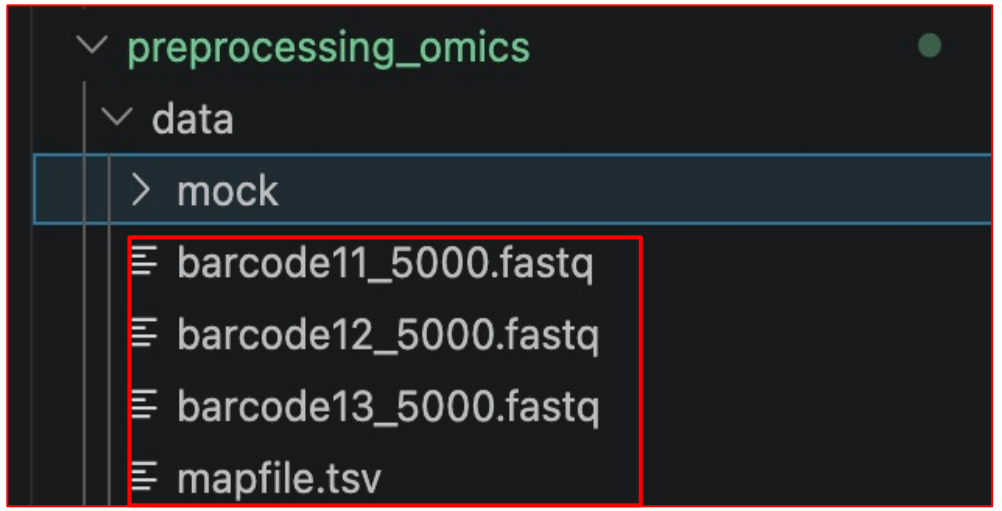
QSCORE	bray_curtis	sum_abs_diff	mse	rmse	mae	VP	FP	FN	VN	recall	FDR	F1
Q07	0,86	28,74	20,46	4,52	3,59	8	0	0	1	1	0	1
Q12	0,86	28	20,5	4,53	3,5	8	0	0	1	1	0	1
Q14	0,85	29,3	18,37	**4,29**	**3,66**	8	0	0	1	1	0	1
Q15	0,86	28,99	17,4	4,17	3,62	8	0	0	1	1	0	1
Q16	0,79	42,91	50,97	7,14	5,36	7	0	1	1	0,875	0	0,933333333
Q17	0,41	113,53	301,66	17,37	14,19	3	1	5	0	0,375	0,25	0,5

- **Recall** = capacité à retrouver les vrais positifs (sensibilité): **Ne pas rater les vrais**
- **FDR** = proportion de faux parmi ce que tu as détecté: **Eviter les faux**
- **F1-score** : équilibre entre détection et précision : **Equilibre global**

Mes données de séquençage des échantillons d'intérêt

On les met dans dossier data

mapfile.tsv aussi si possible



mapefile.tsv pour les échantillons

Mon mapefile.tsv pour mes samples est ok car :

Qscore choisit est 12 → et je respecte le nommage nomsample_12

```
rocessing_omics/data$ more mapfile.tsv
sample_name      code    year    sampleType
barcode11_5000_Q12       F1a     2023    effluent_entree
barcode12_5000_Q12       F2a     2023    actived_sludge
barcode13_5000_Q12       F3a     2023    output_effluent
```

NB : les noms de colonnes n'ont aucune importance MAIS pas de noms avec espace etc

Config.sh file for running analysis of your samples

```
--- Options d'exécution ---
RUN_MOCK=false # désactiver le traitement des mock
RUN_CUTADAPT=true # suppression des primers
RUN_SEQKIT=true # filtrage de qualité
RUN_ADAPT=true # filtrage des adaptateurs
RUN_CHIM=true # filtrage des chimères
RUN_EMU=true # classification EMU

# --- Qualité et filtrage ---
QUAL=12 # score minimum Qscore identifié via les resultats mock
CPU=20 # nombre de threads
MINLEN=1400 # longueur minimale des reads
MAXLEN=1650 # longueur maximale des reads
QUAL_LIST_DEFAULT=() # liste de Qscores par défaut pour mock : a laisser vide
QUAL_LIST=() # liste actuelle de Qscores : a laisser vide
EMUtype="map-ont" # type EMU par défaut

# --- Répertoires de sortie ---
OUTPUT_DIR="./RESULTS_Samples"
```

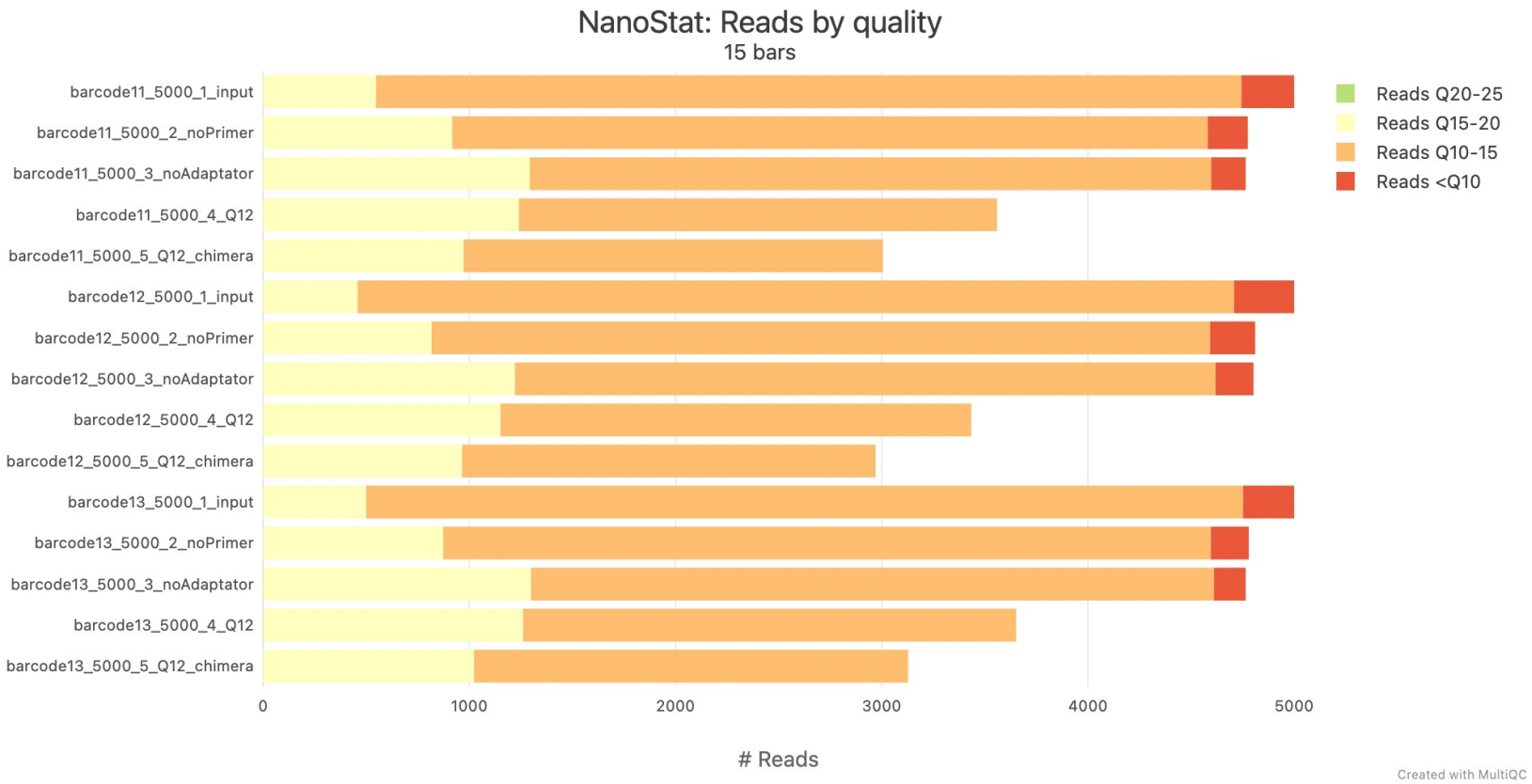
Résultats du préprocessing

Summary Statistics (FASTQ)

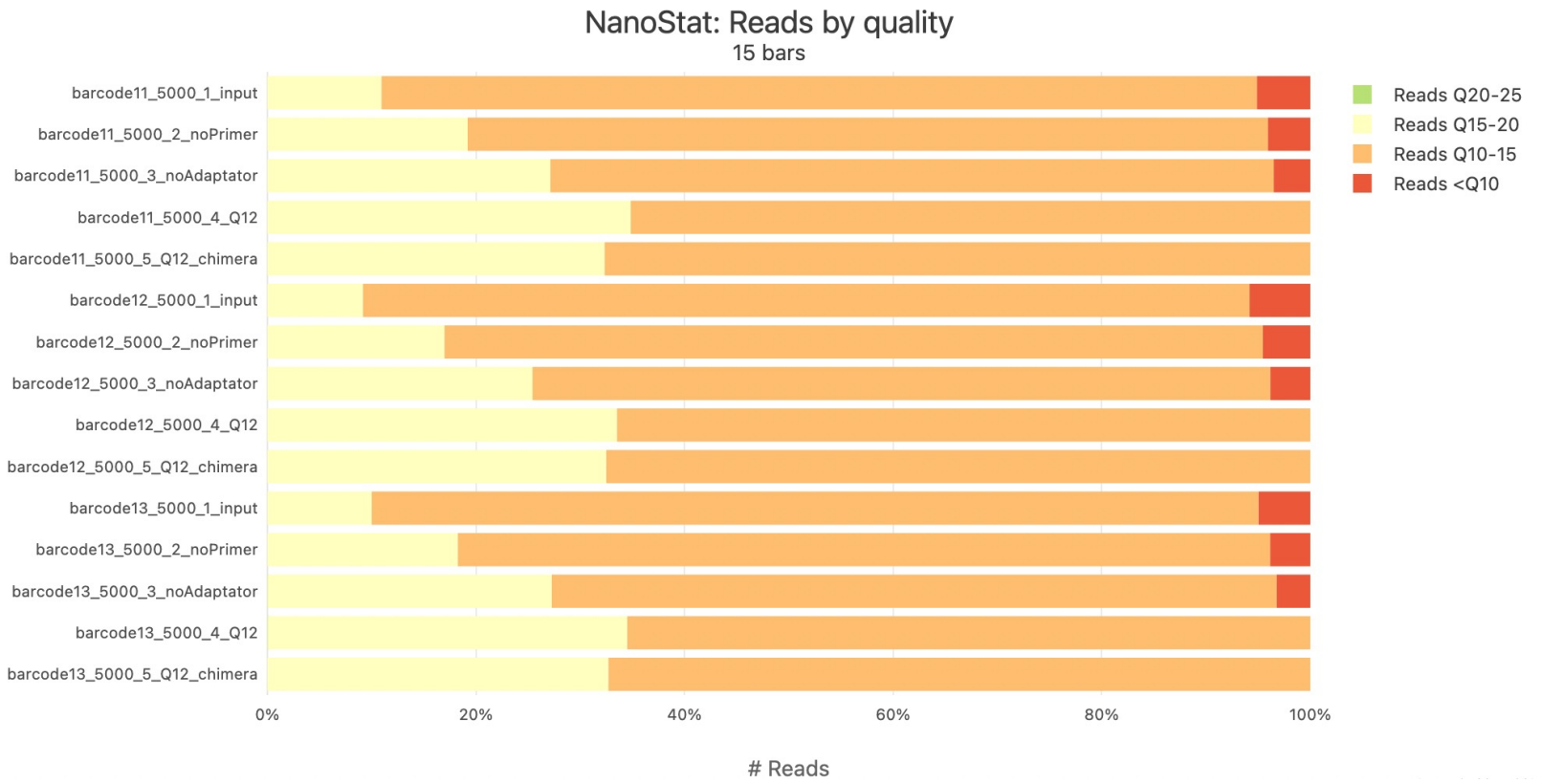
Copy table | Configure columns | Scatter plot | Violin plot | Export as CSV... | Showing 15/15 rows and 5/7 columns. | Summarize table

Sample Name ▲	Median length	Read N50	Median Qual	# Reads (K)	Total Bases (Mb)
barcode11_5000_1_input	1591 bp	1593 bp	13.2	5.0 K	7.8 Mb
barcode11_5000_2_noPrimer	1518 bp	1521 bp	13.5	4.8 K	7.1 Mb
barcode11_5000_3_noAdaptator	1461 bp	1462 bp	13.9	4.8 K	6.8 Mb
barcode11_5000_4_Q12	1465 bp	1466 bp	14.4	3.6 K	5.2 Mb
barcode11_5000_5_Q12_chimera	1465 bp	1466 bp	14.3	3.0 K	4.4 Mb
barcode12_5000_1_input	1578 bp	1579 bp	13.1	5.0 K	7.7 Mb
barcode12_5000_2_noPrimer	1499 bp	1501 bp	13.4	4.8 K	7.1 Mb
barcode12_5000_3_noAdaptator	1449 bp	1451 bp	13.8	4.8 K	6.8 Mb
barcode12_5000_4_Q12	1455 bp	1456 bp	14.4	3.4 K	5.0 Mb
barcode12_5000_5_Q12_chimera	1455 bp	1456 bp	14.3	3.0 K	4.3 Mb
barcode13_5000_1_input	1595 bp	1595 bp	13.2	5.0 K	7.8 Mb
barcode13_5000_2_noPrimer	1514 bp	1518 bp	13.5	4.8 K	7.2 Mb
barcode13_5000_3_noAdaptator	1465 bp	1465 bp	13.9	4.8 K	6.9 Mb
barcode13_5000_4_Q12	1467 bp	1467 bp	14.4	3.7 K	5.4 Mb
barcode13_5000_5_Q12_chimera	1467 bp	1467 bp	14.3	3.1 K	4.6 Mb

Résultats du préprocessing



Résultats du préprocessing



Fichier le plus important : physeq.Rdata (dans Results_Samples)

De la PCA (explorer) vers la RDA (expliquer)

PCA Explore la structure des communautés → Comment les échantillons **se regroupent**
Espèces définissent les gradients

Envfit montre quelles **variables sont liées** à cette organisation → variables **corrélées** aux gradients PC1/PC2

PERMANOVA teste si **ces variables ont vraiment un effet** → **R^2 (variance associée)**
Effet conditionnel
Les R^2 PERMANOVA ≠ fractions strictes

RDA explique les variations abondances des taxons en fonction des variables env → **Structure des communautés** expliquée par les **gradients env**

Le partitionnement de variance décompose, **quantifie** la **contribution des variables env** → **QUI** explique **QUOI** :
unique, partagé, résiduel