Introduction aux méthodes de séquençage, à la bioinformatique et aux statistiques

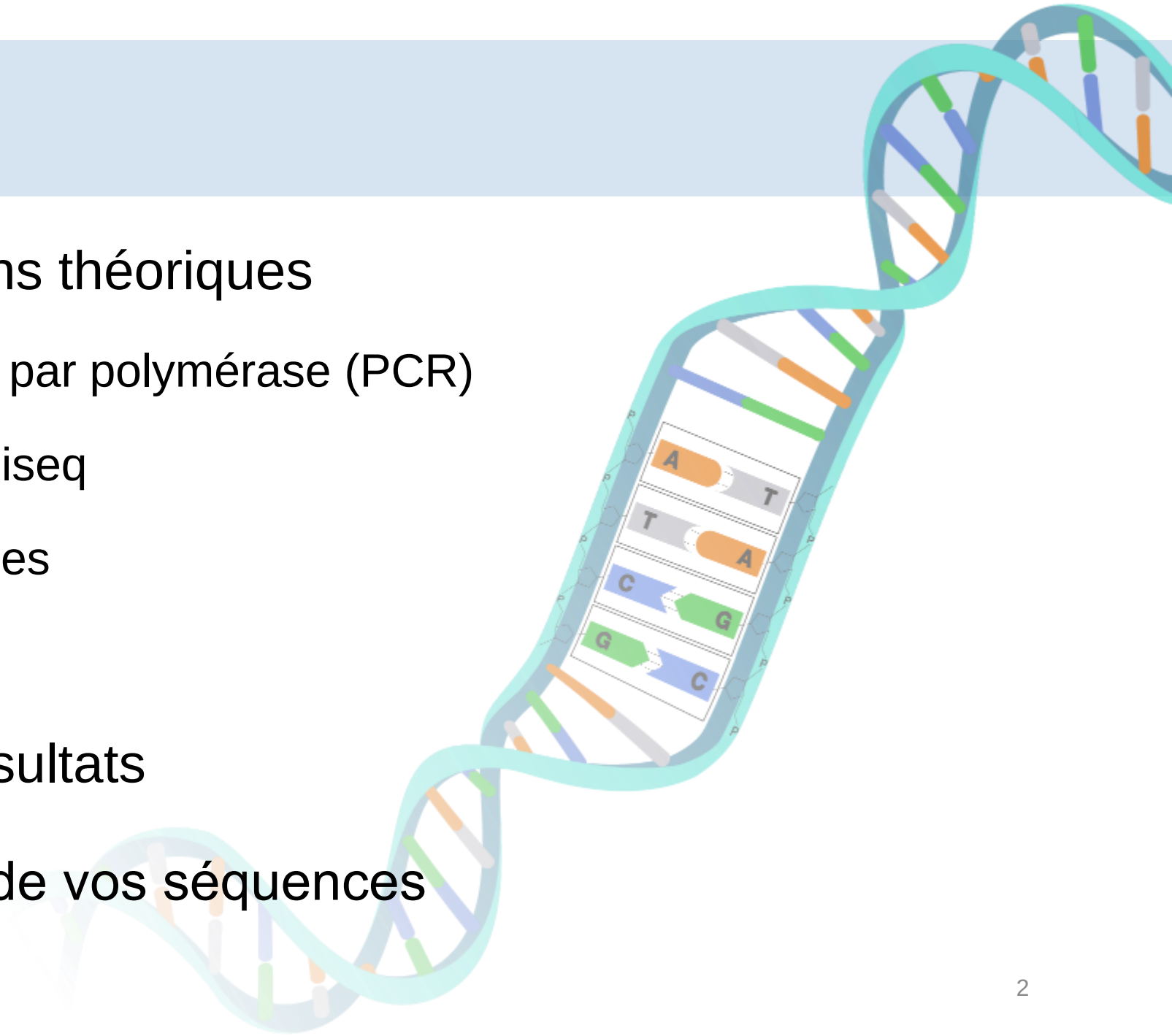
BIO1410 Hiver 2025

PK-S1535

Microbiologie Environnementale

Plan de la séance

- Présentation des notions théoriques
  - Amplification en chaîne par polymérase (PCR)
  - Séquençage Illumina Miseq
  - Analyse bio-informatiques
  - BLAST
- Présentation de vos résultats
- Analyse BLAST d’une de vos séquences



Notions théoriques

- **Nous avons…**
  - Des échantillons composés de plusieurs microorganismes différents

- **La problématique…**
  - Comment pouvons-nous identifier simultanément le plus de microorganismes tout en étant capable de distinguer les différents taxons?
- **La solution…**
  - Le séquençage de régions spécifiques de l'ADN

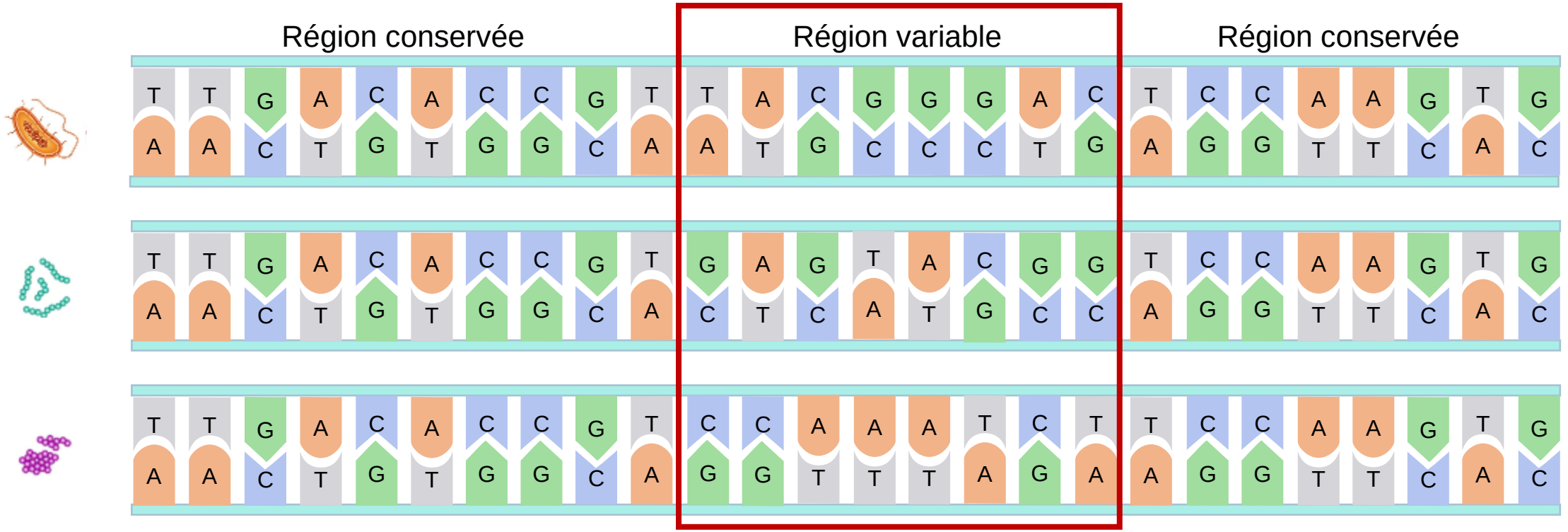
Notions théoriques

Certaines séquences d'ADN sont composées de :

- **régions hyper-conservées**
- **régions variables**

Permet à la fois de :

- **cibler plusieurs organismes**
- **distinguer ces organismes les un des autres**



Gène ADNr 16S et 18S

Séquence d'ADN (gène ADNr) codant pour un ARN ribosomique (ARNr)

Procaryotes

Archées et Bactéries

ARN ribosomique 16S

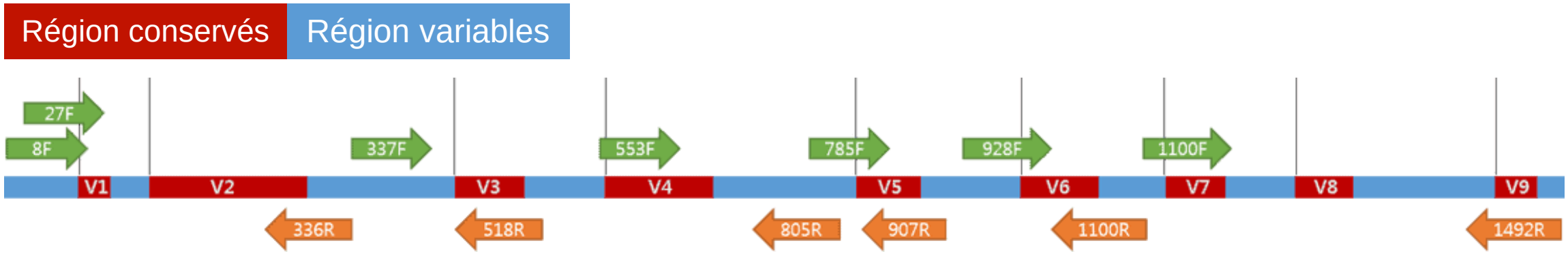
Gène : ARNr 16S

Eucaryotes

Animaux, végétaux, protistes, champignons

ARN ribosomique 18S

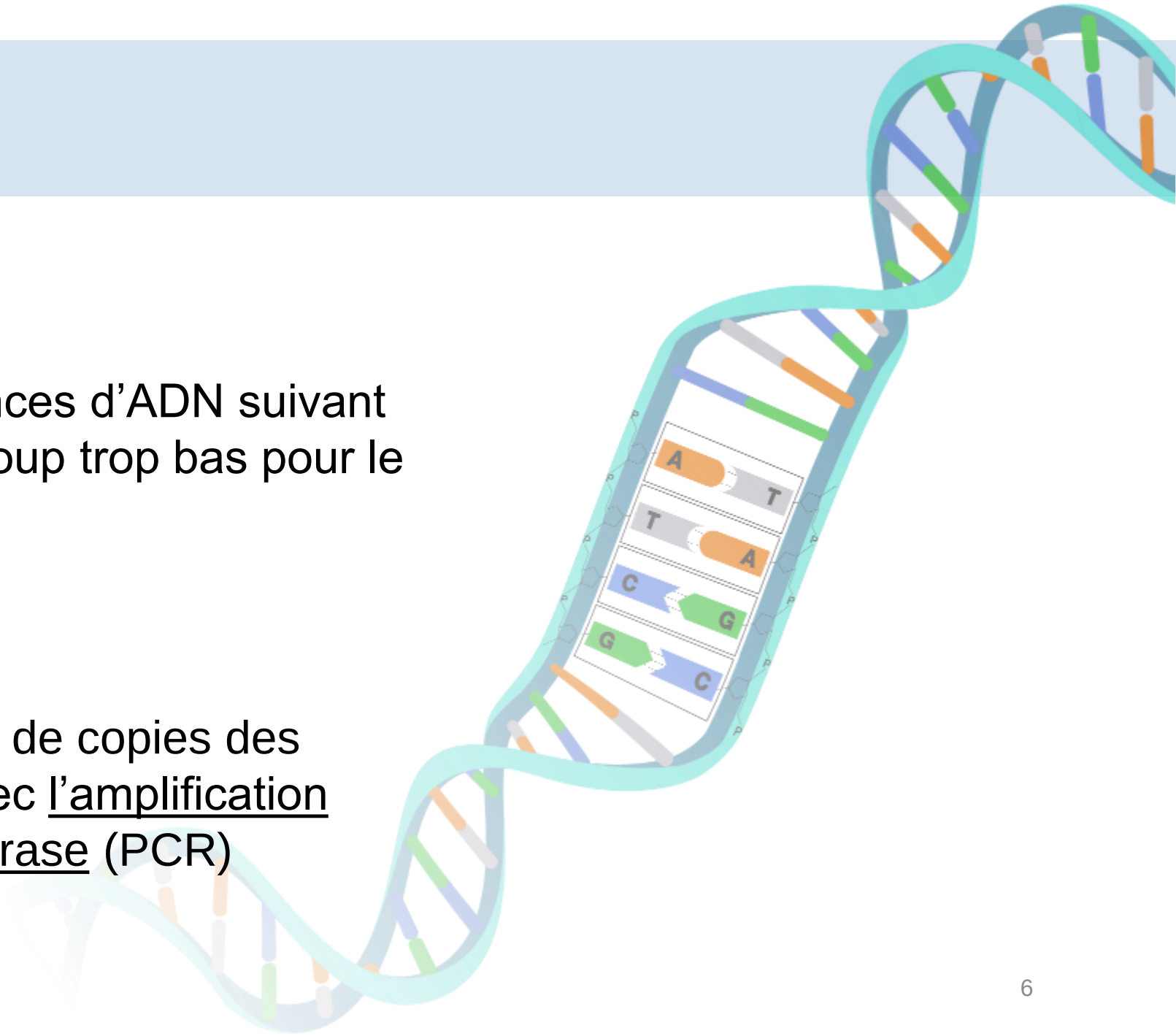
Gène : ARNr 18S



Représentation des régions conservées et variables du gène procaryote ADNr 16S

Notions théoriques

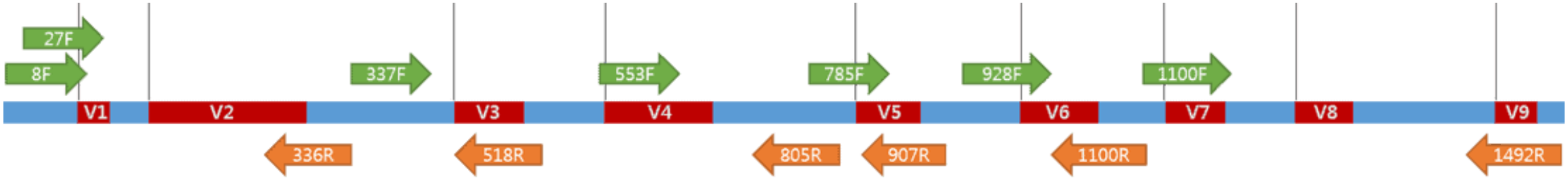
- **La problématique**
  - Le nombre de séquences d'ADN suivant l'extraction est beaucoup trop bas pour le séquençage
- **La solution**
  - Augmenter le nombre de copies des séquences d'ADN avec l'amplification en chaîne par polymérase (PCR)



Amplification en chaîne par polymérase

- Objectif
  - Augmenter le nombre de séquences d'ADN d'intérêt
- Étapes
  - Dénaturation
  - Hybridation
  - Élongation
- Région et amorces
  - V5-V6
  - B799F (AACMGGATTAGATACCCKG) et B1115R (AGGGTTGCGCTCGTTG)

Thermocycleur servant à la PCR

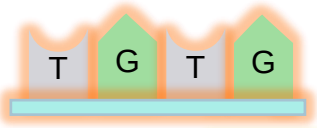


Amplification en chaîne par polymérase

Solution

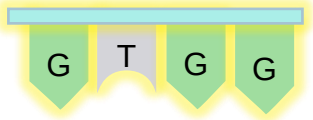
Polymérase

Amorces

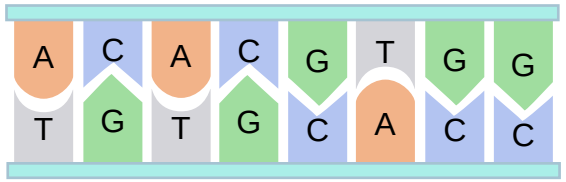


Forward

Reverse



ADN extrait à amplifier



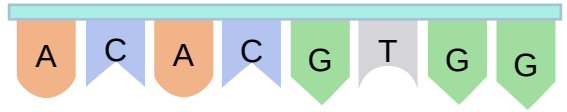
Nucléotides (dNTP)

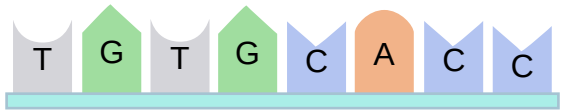
Les étapes 1 à 3 sont répétées de 35 à 40 fois et pour obtenir ~ 1 million de copies d'une séquence

1. Dénaturation

95°C

Séparation de l'ADN en deux brins

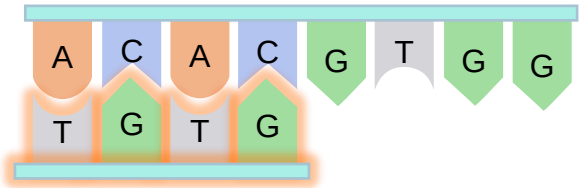




2. Hybridation

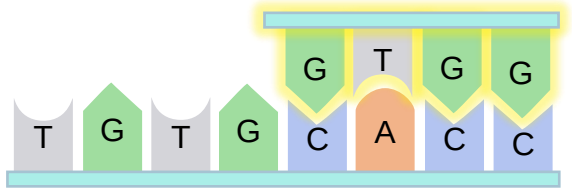
50-65°C

Attachement des amorces



Amorce *Forward*

Amorce *Reverse*

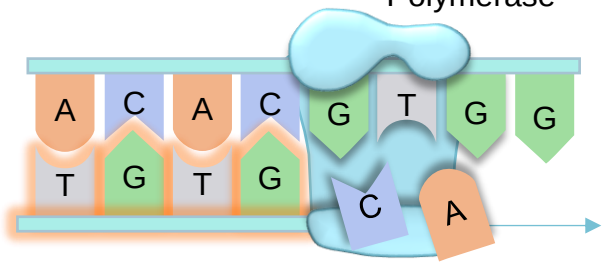


3. Élongation

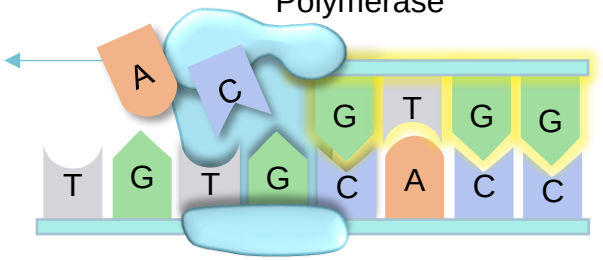
75°C

La polymérase vient synthétiser le fragment complémentaire

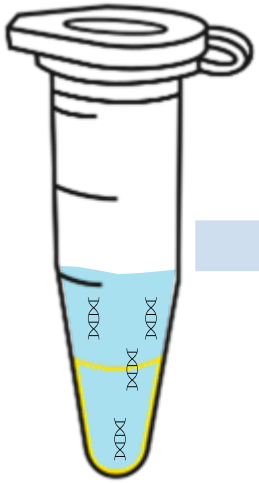
Polymérase



Polymérase

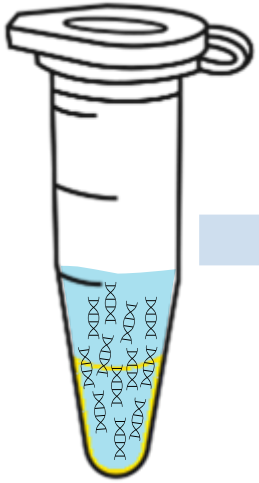


Synthèse



ADN extrait
Peu d'ADN

Amplification
par PCR



Produit de la PCR
Beaucoup d'ADN

Lecture des séquences
Illumina Miseq

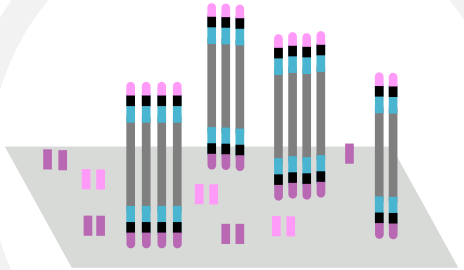
Analyses
bio-informatiques

Séquençage des amplicons

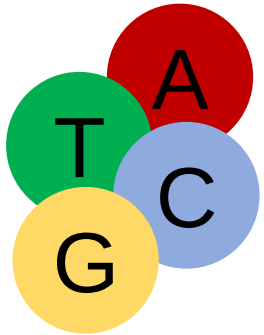
- Objectif
  - Déterminer l'ordre d'enchaînement des nucléotides

Principales étapes

Préparation des libraires



Génération d’amplifiats



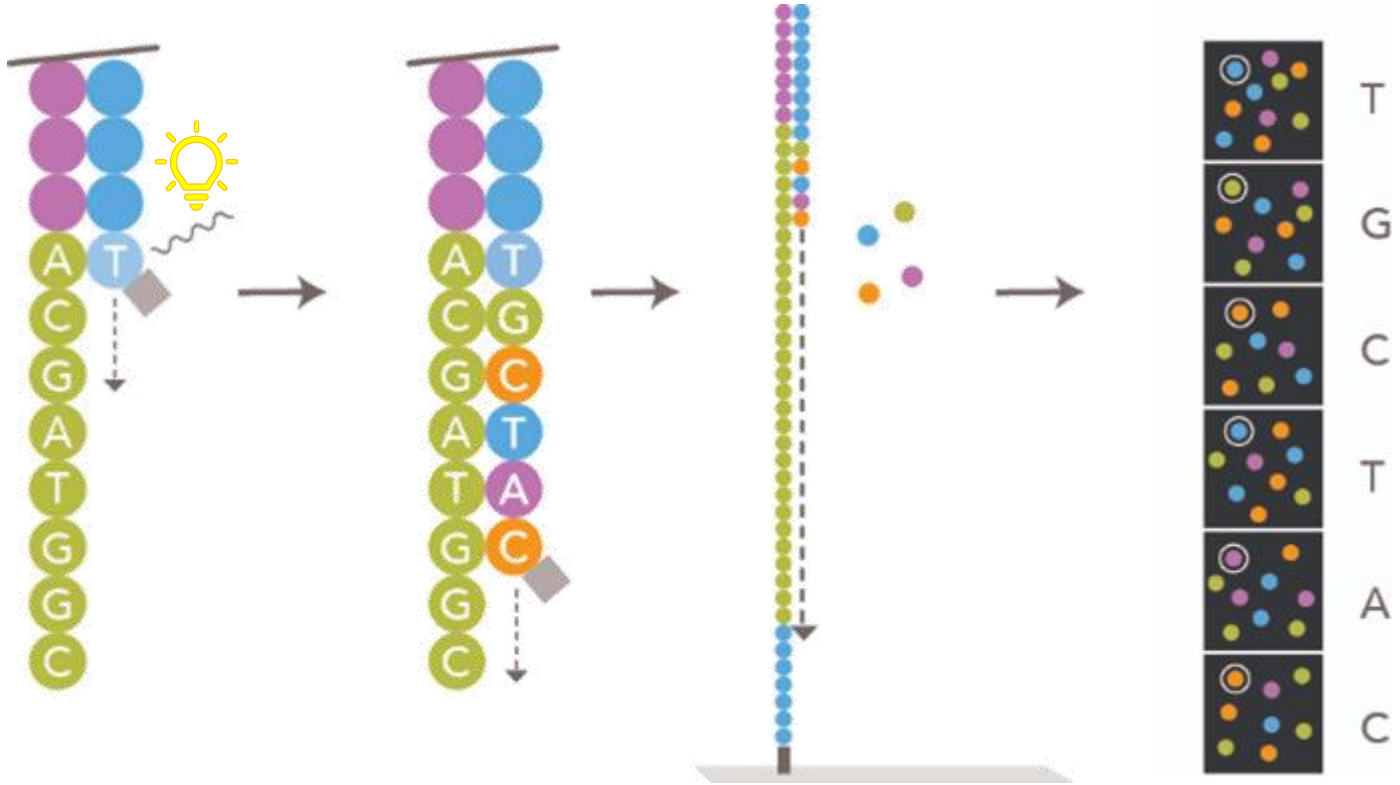
Séquençage

Description des étapes : https://www.youtube.com/watch?v=fCd6B5HRaZ8&ab_channel=Illumina

Séquençage

Les nucléotides fluorescents s'hybrident au brin d'ADN :

Une hybridation par cycle, le nombre de cycle est déterminé par la longueur du fragment



Les nucléotides incorporés sont identifiés selon leur signaux fluorescents. Les clusters s'illuminent simultanément permettant la lecture

Le processus entier est répété pour le brin reverse

Produits du séquençage

Fichier texte (fastq / fasta) avec l'ensemble des séquences d'acides nucléiques identifiés par le séquenceur

Exemple :

Brin *Forward* (R1)

ACTCGCTATTCGACTAGAACCGGATTAGATACCCTGGTAGTCC
ACGCCGTAAACGGTGGACGCTGGATGTGGGGCCCATTCCAC
GGGTTCTGTGTCGGAGCTAACGCGTTAAGCGTCCCGCCTGG
GGAGTACGGCCGCAAGGCTAAAACTCAAAGAAATTGACGGG
GGCCCGCACAAGCGGCGGAGCATGCGGATTAATTCGATGCA
ACGCGAAGAACCTTACCTGGGCTTGACATGTGCCTGACGAC
TGCAGAGATGTGGTTTCCT

Brin *Reverse* (R2)

ACTCGCTARTCGACTAGAGGGTTGCGCTCGTTGCGGGACTT
AACCCAACATCTCACGACACGAGCTGACGACGACCATGCAC
CACCTGTGAACCTGCCCCGAAAGGAAACCACATCTCTGCAG
TCGTCAGGCACATGTCAAGCCCAGGTAAGGTTCTTCGCGTT
GCATCGAATTAATCCGCATGCTCCGCCGCTTGTGCGGGCCC
CCGTCAATTTCTTTGAGTTTTAGCCTTGCGGCCGTACTCCCC
AGGCGGGACGCTTAACGCGTT

Analyses bio-informatiques

- Objectif
  - Identifier les microorganismes échantillonnés
    et répondre aux hypothèses à l'aide de méthodes statistiques
- Étapes du traitement des séquences
  - Filtrer et rogner les séquences en fonction de la qualité
  - Retirer les erreurs au niveau des nucléotides
  - Unifier les brins *Forward* et *Reverse*
  - Retirer les séquences chimères
  - Générer une table de variants de séquence d'amplicon (ASV)
  - Assigner une taxonomie aux ASVs (base de données Silva)

Analyses bio-informatiques

Brin *Forward* (R1)

ACTCGCTATTCGACTAGAACCGGATTAGATACCCTGGTAGTCCACGCCGTAAACGGTGGACGCTGGATGTGGGGCCCATTCCACG
GGTTCTGTGTCGGAGCTAACGCGTTAAGCGTCCCGCCTGGGGAGTACGGCCGCAAGGCTAAAACTCAAAGAAATTGA
CGGGGGCCCGCACAAGCGGCGGAGCATGCGGATTAATTCGATGCAACGCGAAGAACCTTACCTGGGCTTGACATGTGCCTGACGACTGCAGAGATGTGGTTTCCT

Brin *Reverse* (R2)

ACTCGCTARTCGACTAGAGGGTTGCGCTCGTTGCGGGACTTAACCCAACATCTCACGACACGAGCTGACGACGACCATGCACCACCTGTGAACCT
GCCCCGAAAGGAAACCACATCTCTGCAGTCGTCAGGCACATGTCAAGCCCAGGTAAGGTTCTTCGCGTTGCATCGAATTAATCCGCATGCTCCGCCGCTTGTGCGGGCCCCCGTCAATTTCTTTGAGTTTTAGCCTTGCGGCCGTACTCCCCAGGCGGGACGCTTAACGCGTT

Retirer les amorces F et R et corriger les erreurs (étapes 1 à 2)

CGACTAGAACCGGATTAGATACCCTGGTAGTCCTCGCCGTAAACGGTGGACGCTGGATGTGGGGCCCATTCCACG
GGTTCTGTGTCGGAGCTAACGCGTTAAGCGTCCCGCCTGGGGAGTACGGCCGCAAGGCTAAAACTCAAAGAAATTGA
CGGGGGCCCGCACAAGCGGCGGAGCATGCGGATTAATTCGATGCAACGCGAAGAACCTTACCTGGGCTTGACATGTGCCTGACGACTGCAGAGATGTGGTTTCCT

ACTCGCTARTCGACTAGAGGGTTGCGCTCGTTGCGGGACTTAACCCAACATCTCACGACACGAGCTGACGACGACCATGCACCACCTGTGAACCT
GCCCCGAAAGGAAACCACATCTCTGCAGTCGTCAGGCACATGTCAAGCCCAGGTAAGGTTCTTCGCGTTGCATCGAATTAATCCGCATGCTCCGCCGCTTGTGCGGGCCCCCGTCAATTTCTTTGAGTTTTAGCCTTGCGGCCGTACTCCCCAGGCGGGAC

Unifier les brins Forward et Reverse (étape 3)

Zones de chevauchement

CGACTAGAACCGGATTAGATACCCTGGTAGTCCTCGCCGTAAACGGTGGACGCTGGATGTGGGGCCCATTCCACGCTCG
ACGCTGGATGTGGGGCCCATTCCACGCTCGTTGCGGGACTTAACCCAACACTCACGACACGAGCTGACGACGACCATGCACCACCTGTGAACC
GGTTCTGTGTCGGAGCTAACGCGTTAAGCGTCCCGCCTGGGGAGTACGGCCGCAAGGCTAAAACTCAAAGAAATTGAGG
CCGCAAGGCTAAAACTCAAAGAAATTGAGGAAACCACATCTCTGCAGTCGTCAGGCACATGTCAAGCCCAGGTAAGGTTCTTCGCGTTGCATC
CGGGGGCCCGCACAAGCGGCGGAGCATGCGGATTAATTCGATGCAACGCGAAGAACCTTACCTGGGCTTGACATGTGCC
GAAGAACTTACCTGGGCTTGACATGTGCCTTTGTGCGGGCCCCCGTCAATTTCTTTGAGTTTTAGCCTTGCGGCCGTACTCCCCAGGCGGGAC

Retirer les chimères (étape 4)

CGACTAGAACCGGATTAGATACCCTGGTAGTCCTCGCCGTAAACGGTGGACGCTGGATGTGGGG[X]TCCACGCTCGCTGGGGAGTACGGCCGCAAGGCTAAAACTCAAAGAAATTGAGG

Analyses bio-informatiques

Générer une table d'ASVs (étape 5)

CGACTAGAACCGGATTAGATACCCTGGTAGTCCTCGCCGTAAACGGTGGACGCTGGATGTGGGGCCCATTCCACGC
CGACTAGAACCGGATTAGATACCCTGGTAGTCCTCGCCGTAAACGGTGGACGCTGGATGTGGGGCCCATTCCACGCTCGTTGCGGGACTTAACCCAACACTCACGACACGAGCTGACGACGACCATGCACCACCTGTGAACC
CGACTAGAACCGGATTAGATACCCTGGTAGTCCTCGCCGTAAACGGTGGACGCTGGATGTGGGGCCCATTCCACGCTCGTTGCGGGACTTAACCCAACACTCACGACACGAGCTGACGACGACCATGCACCACCTGTGAACC
GGTTCTGTGTCGGAGCTAACGCGTTAAGCGTCCCGCCTGGGGAGTACGGCCGCAAGGCTAAAACTCAAAGAAATTGAGGAAACCACATCTCTGCAG
GTTCTGTGTCGGAGCTAACGCGTTAAGCGTCCCGCCTGGGGAGTACGGCCGCAAGGCTAAAACTCAAAGAAATTGAGGAAACCACATCTCTGCAG
CGGATTAATTCGATGCAACGCGAAGAACCTTACCTGGGCTTGACATGTGCCTTGTGCGGGCCCCCGTCAATTTCTTTGAGTTTTAGCCTTG
CGGATTAATTCGATGCAACGCGAAGAACCTTACCTGGGCTTGACATGTGCCTTGTGCGGGCCCCCGTCAATTTCTTTGAGTTTTAGCCTT

- Grouper les séquences identiques ensemble sous un même ASV
- On obtient donc des groupes de séquences comparables a des espèces
- Nombre de séquences correspondent à l'abondance de ce taxon

ASV	Séquence	Nombre
ASV1	CGACTAGAACCGGATTAGATACCCTGGTAGTCCTCGCCGTAAACGGTGGACGCTGGATGTGGGGCCCATTCCACGCTCGTTGCGGGACTTAACCCAACACTCACGACACGAGCTGACGACGACCATGCACCACCTGTGAACC	3
ASV2	CGGATTAATTCGATGCAACGCGAAGAACCTTACCTGGGCTTGACATGTGCCTTGTGCGGGCCCCCGTCAATTTCTTTGAGTTTTAGCCTTG	2
ASV3	GGTTCTGTGTCGGAGCTAACGCGTTAAGCGTCCCGCCTGGGGAGTACGGCCGCAAGGCTAAAACTCAAAGAAATTGAGGAAACCACATCTCTGCAG	2

Assigner une taxonomie aux ASVs (étape 6)

- Classification par comparaison des séquences à une base de données

silva
high quality ribosomal RNA databases

Basic Local Alignement Search Tool (BLAST)

- Objectif :
  - Comparer les séquences de nucléotides ou d'acides aminées d'intérêts avec ceux de la base de données et calculer la signifiance statistique du *match*

Choix de la base de données appropriée:

- rRNA Database
  - 16s Ribosomal RNA sequeneces

Type de Blast:

- Blastn: algorithme de base utilisé pour tous types de séquences
- MegaBlast : Plus rapide, mais n'aligne que les séquences identiques à plus de 95%

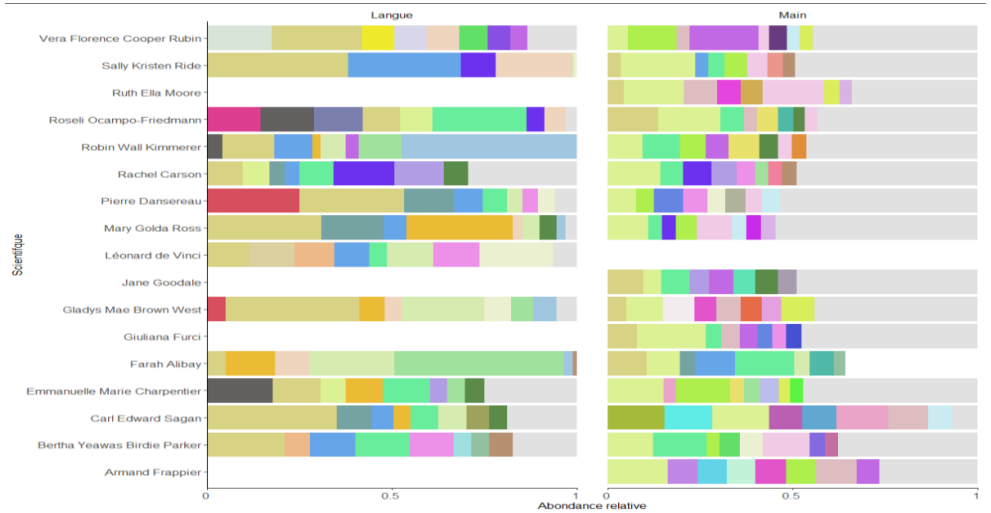
BLAST

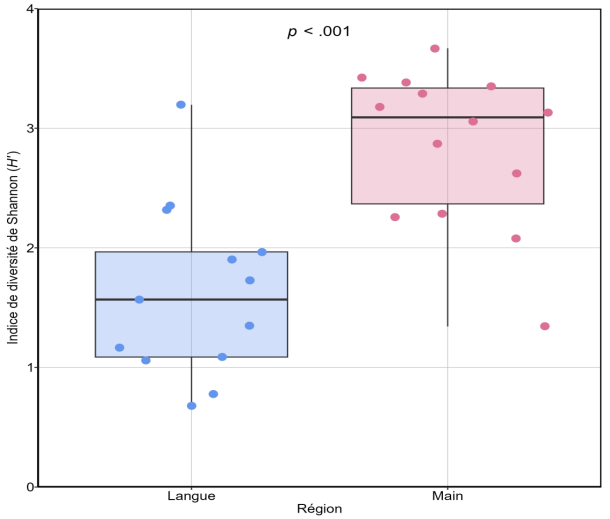
Paramètres à prendre en compte pour l'identification:

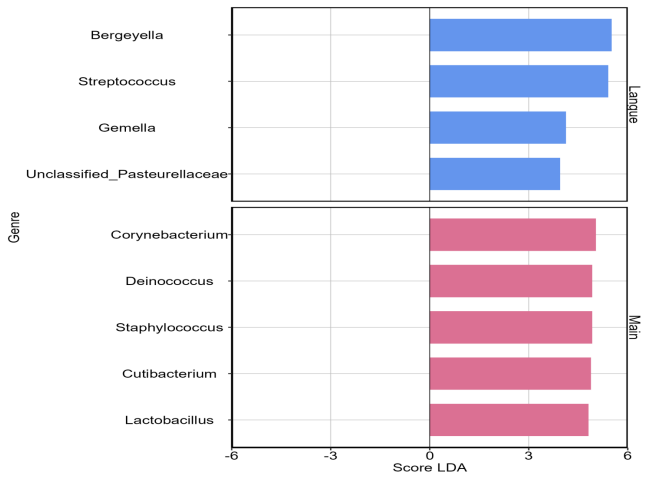
- **Per. Ident** : pourcentage d'identification
- **Query cover** : Fraction de la séquence s'alignant avec notre séquence
- **E value** : Nombre de séquences de qualité similaire qui pourraient être trouvé par chance

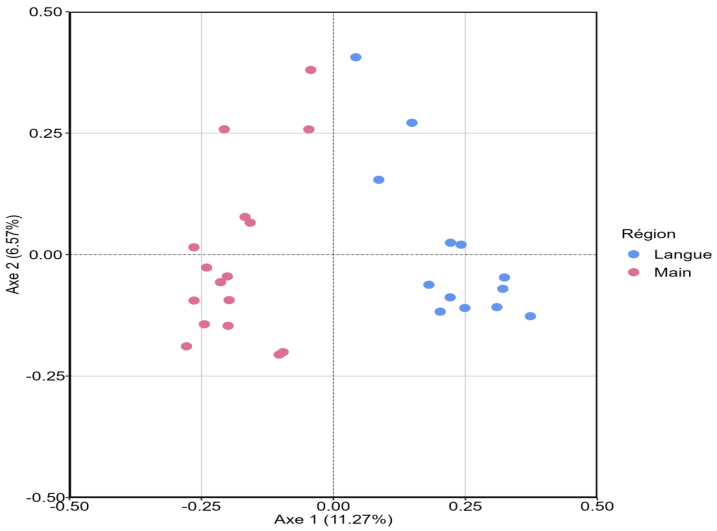
Scientific Name	Max Score	Total Score	Query Cover	**E value**	Per. Ident	Acc. Len
Staphylococcus …	551	3302	100%	1e-152	100.00%	2852412

Vos résultats

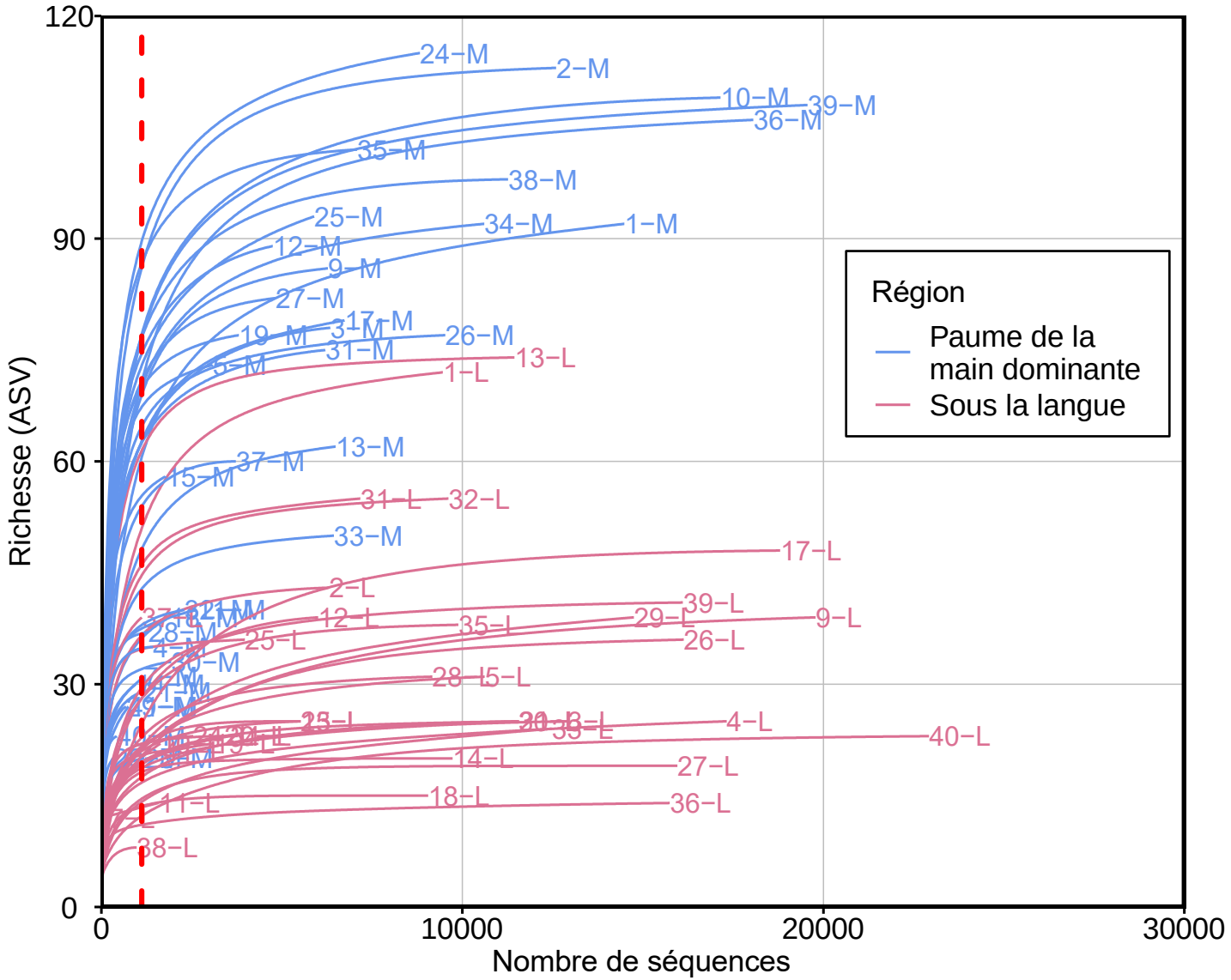








Raréfier



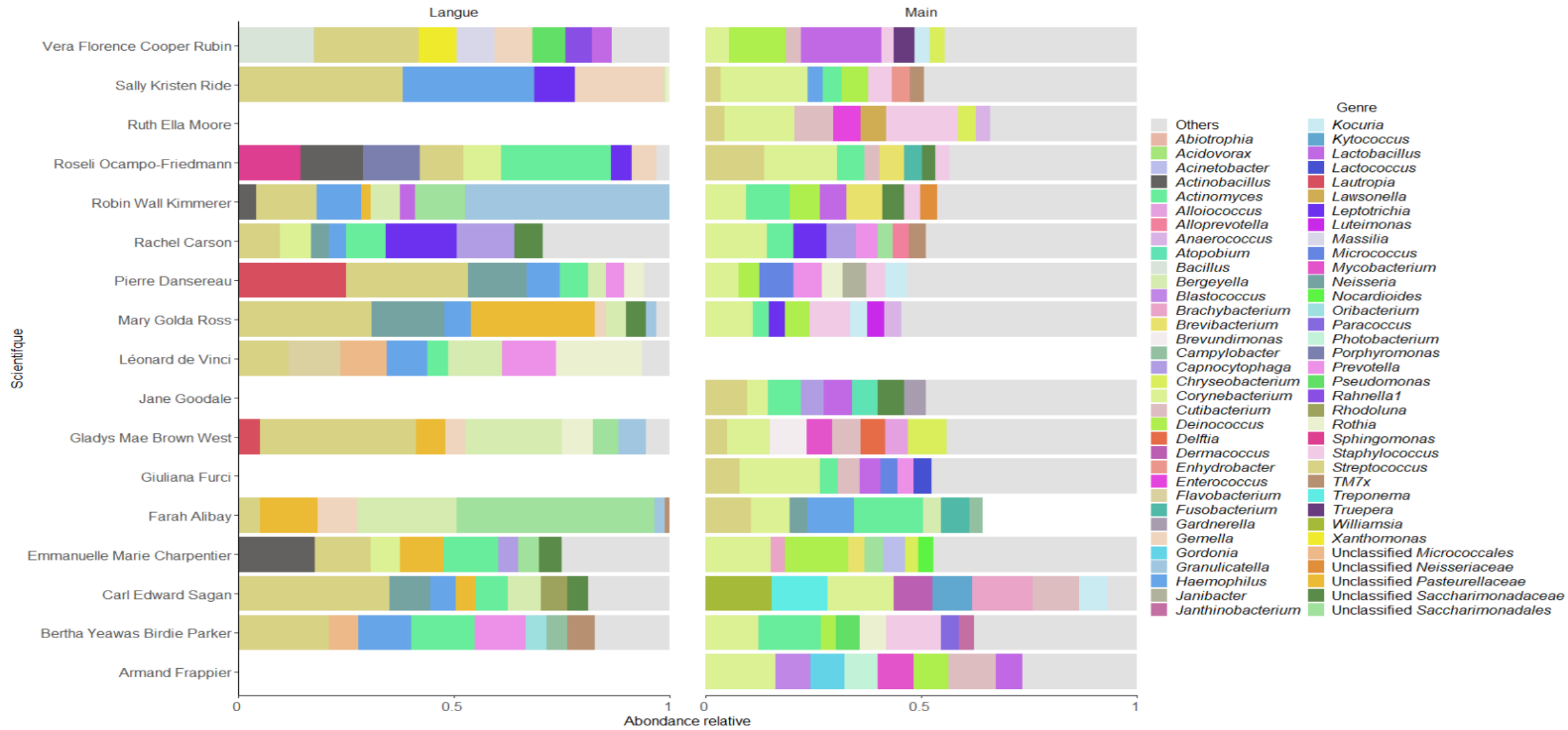
Pourquoi raréfier?

- Contrôler pour l'inéquitabilité dans l'effort de séquençage
- c-à-d contrôler pour la variation dans le nombre de séquences obtenues par échantillon (pour des raisons techniques et non biologiques)

Pour plus d'infos:
Schloss, P. D. (2024). Rarefaction is currently the best approach to control for uneven sequencing effort in amplicon sequence analyses. *Msphere*, e00354-23.

→ Raréfaction à 1118 séquences

Abondance relative des principaux genres

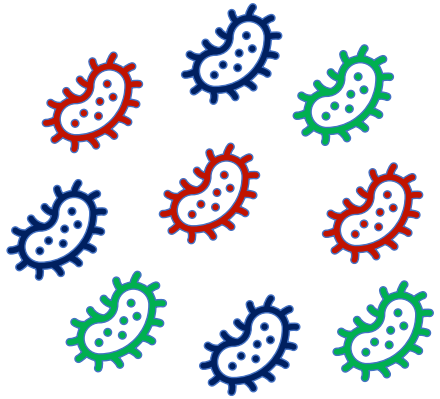


Diversité alpha

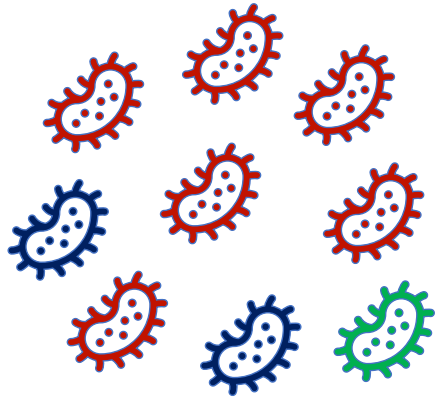
- Qu'est-ce que la diversité alpha?
  - Diversité d'une communauté à l'échelle locale (ex. diversité d'espèces bactériennes sur la paume d'une personne)
- Comment peut-on la mesurer?
  - Richesse
  - Indice de Shannon
  - autres indices (ex. Simpson, Chao, etc.)

Diversité alpha – Indice de Shannon

- Quels sont les deux éléments pris en compte pour le calcul de l'indice de Shannon?
  - Nombre d'unités taxonomiques distinctes (ex. espèces, ASV, etc.)
  - Équitabilité (abondance relative)



Comm. A

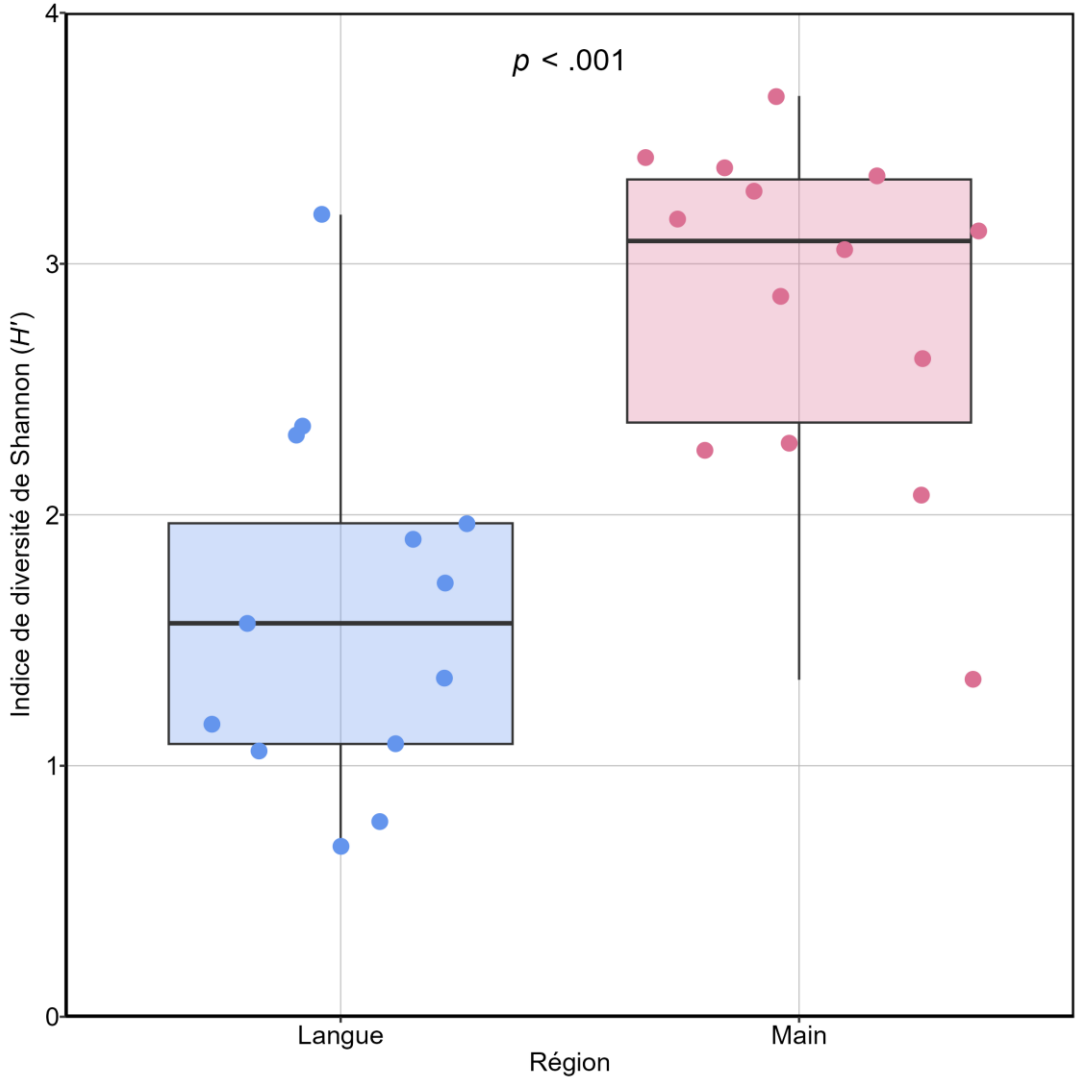


Comm. B

Diversité alpha

- D'après vous, quel habitat (paume ou langue) possède la plus grande diversité alpha (Shannon)? Pourquoi?

Résultats – Indice de Shannon



Mesurer l'indice de diversité de **Shannon**

Utiliser le test de ***t* de Student** pour comparer l'indice de Shannon en fonction de la région échantillonnée

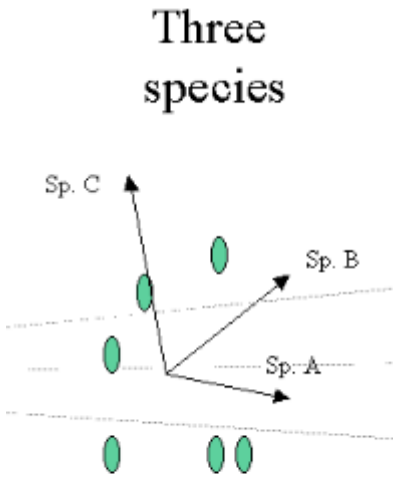
À vous d'interpréter <u>biologiquement</u> ce résultat dans votre rapport!

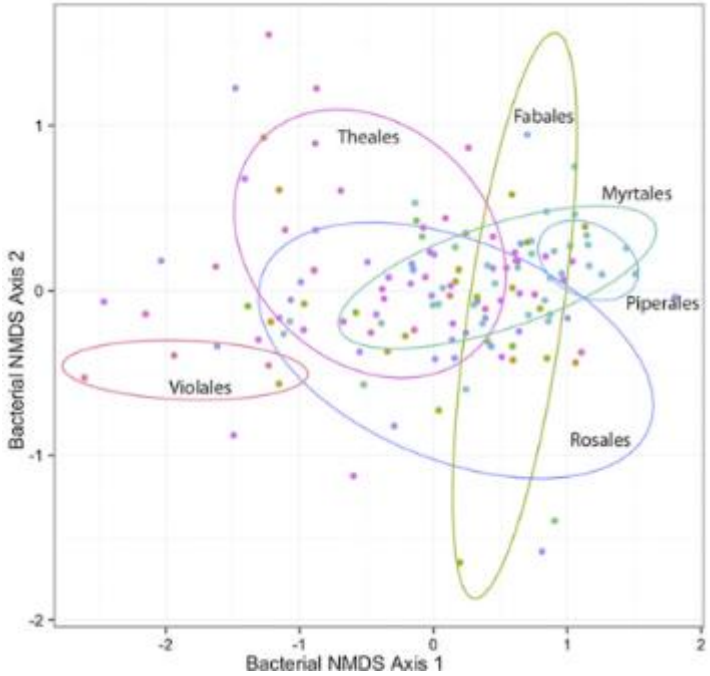
Diversité bêta

- Qu’est-ce que la diversité bêta?
  - Représente la variation dans la composition des espèces (ou des ASV) d’une communauté à l’autre
- À quelles questions écologiques peut-on répondre avec des analyses de diversité bêta?
  - Est-ce que différents types d’habitats sont composés de différentes espèces?
  - Est-ce qu’un habitat est plus hétérogène qu’un autre en termes de composition en espèces?

Diversité bêta – Ordinations

- Permettent de visualiser quels échantillons sont plus semblables ou plus différents en termes de composition en espèces
- Nos échantillons peuvent être positionnés dans un espace multidimensionnel en fonction de leur composition, où chaque axe (ou dimension) représente un taxon (ex. ASV).
- Dans cet espace multidimensionnel, les échantillons situés plus proches l'un de l'autre sont plus similaires dans leur composition en espèces que les échantillons situés plus loin.



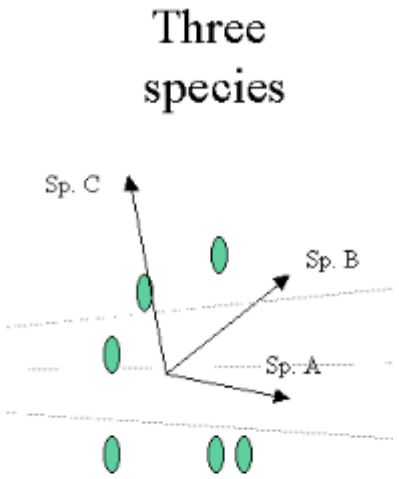


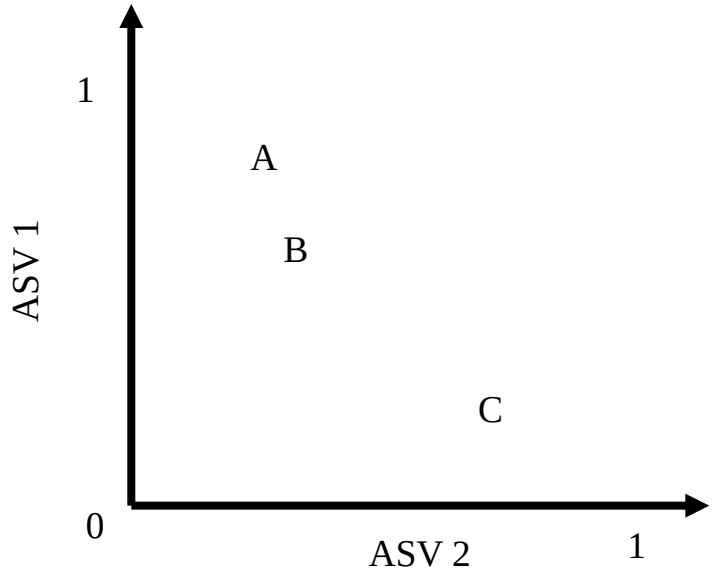
Kembel et al. 2014

Source : Michael W. Palmer; https://ordination.okstate.edu/overview.htm.

Diversité bêta – Ordinations

- Permettent de visualiser quels échantillons sont plus semblables ou plus différents en termes de composition en espèces
- Nos échantillons peuvent être positionnés dans un espace multidimensionnel en fonction de leur composition, où chaque axe (ou dimension) représente un taxon (ex. ASV).
- Dans cet espace multidimensionnel, les échantillons situés plus proches l'un de l'autre sont plus similaires dans leur composition en espèces que les échantillons situés plus loin.

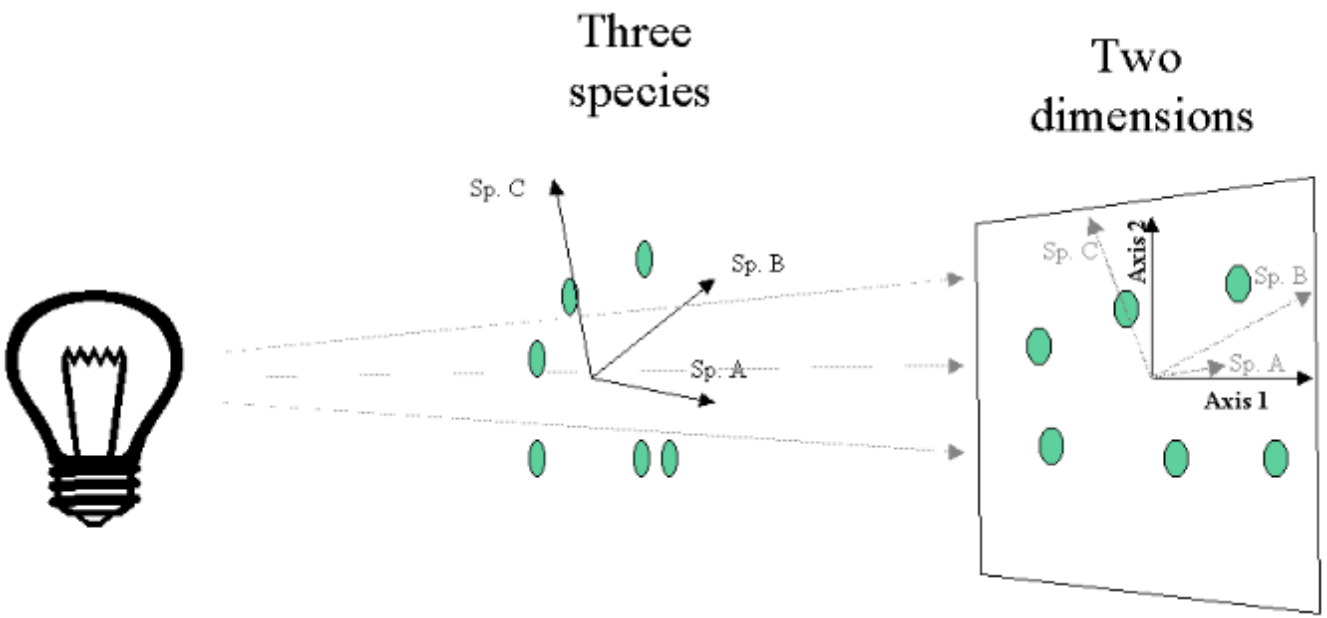




Source : Michael W. Palmer; https://ordination.okstate.edu/overview.htm.

Diversité bêta – Ordinations

- L'ordination trouve de nouveaux axes dans l'espace multidimensionnel, pour visualiser l'étendue de la variance dans nos données compositionnelles (souvent avec les deux premiers nouveaux axes qui représente le plus grand pourcentage de variance).
- Ces nouveaux axes représentent une "combinaison de variables" (et non plus des espèces particulières).

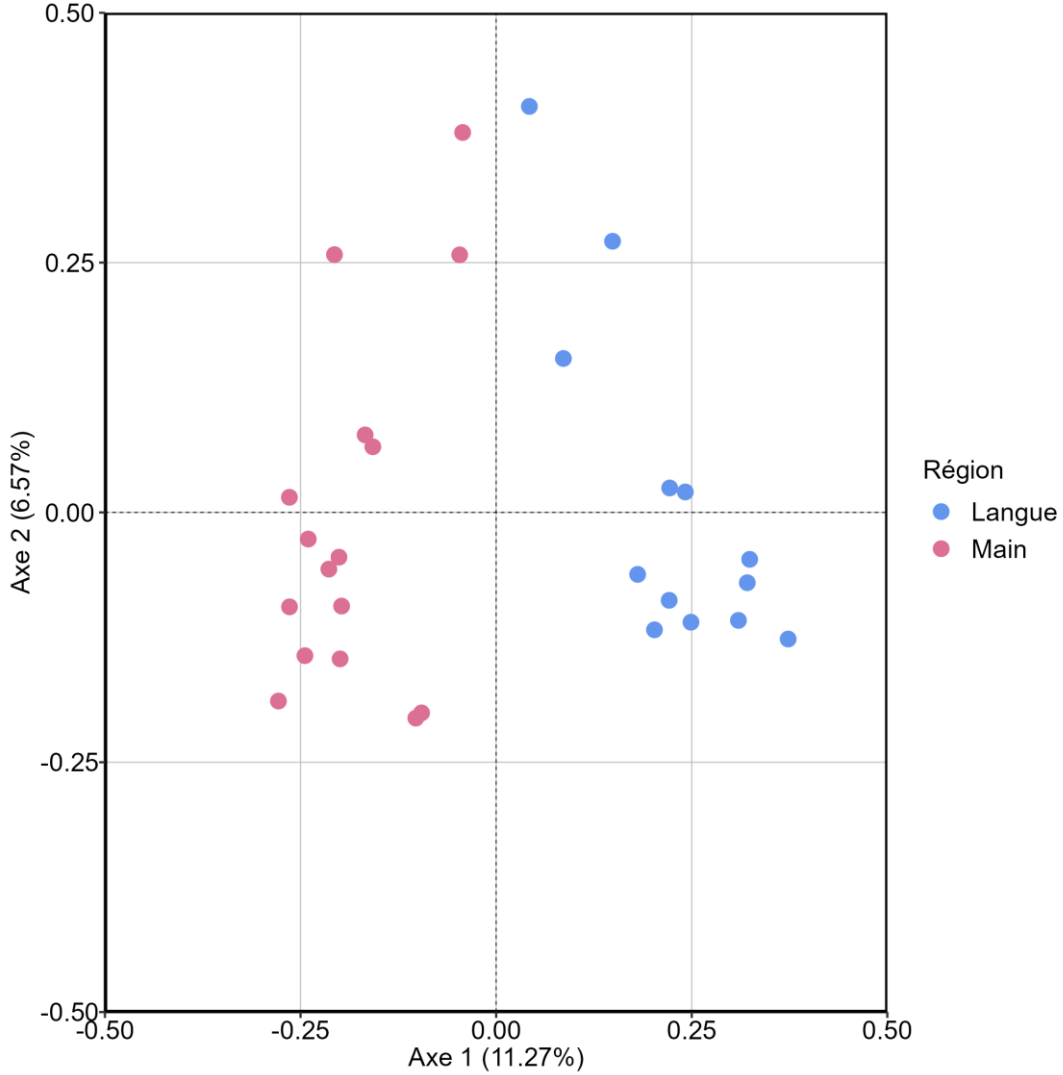


→ L'ordination permet de visualiser l'espace multidimensionnel en 2D (espace réduit) et simplifie l'interprétation des données multivariées.

→ Il existe différentes méthodes d'ordination (ex. PCA, PCoA)

Source : Michael W. Palmer; https://ordination.okstate.edu/overview.htm.

Résultats – Ordination (PCoA)



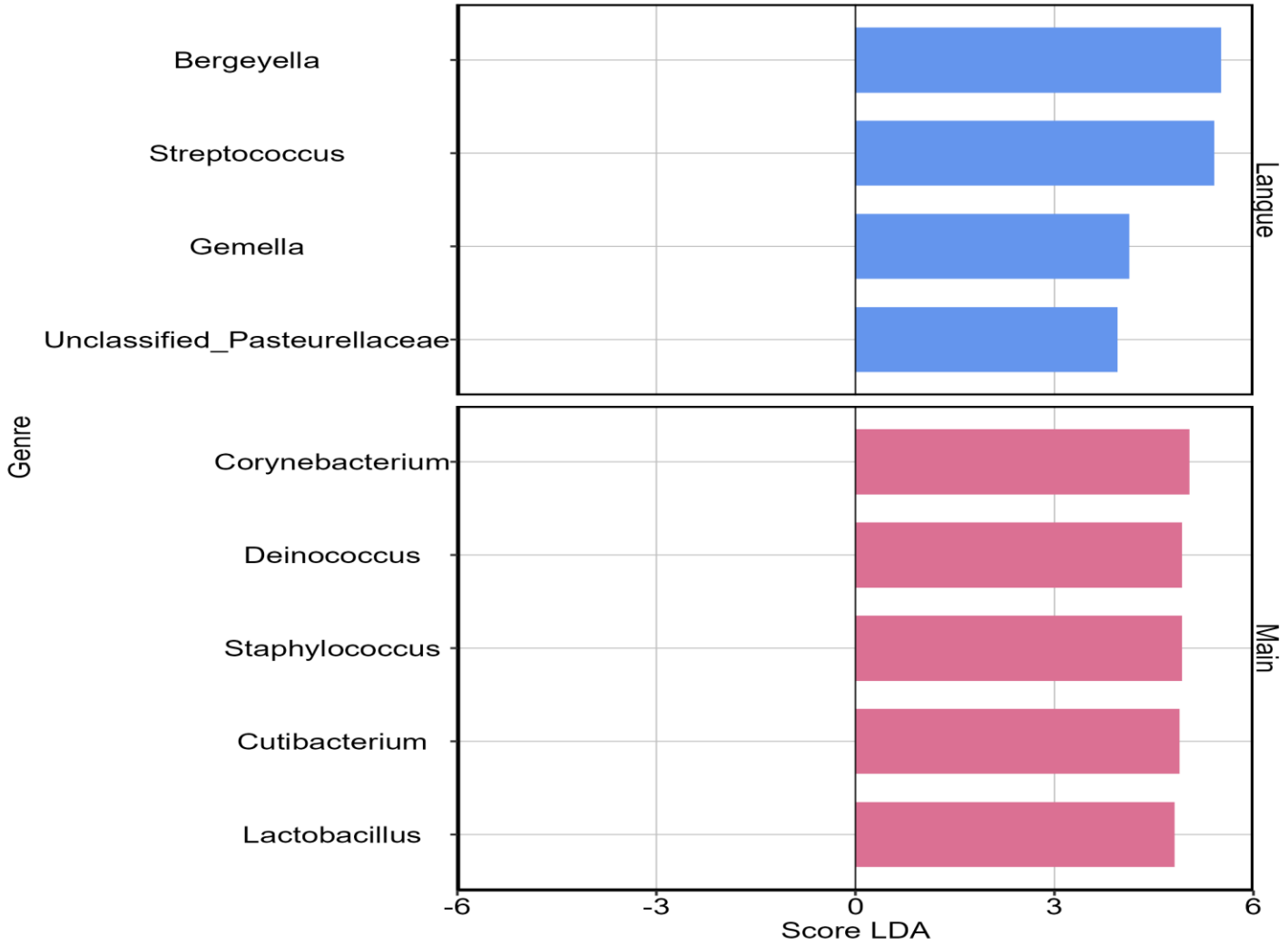
Quelles informations écologiques peut-on tirer de cette ordination?

Quelles sont les <u>raisons biologiques</u> pouvant expliquer les différences observées?

Région
- Paume de la main dominante
- Sous la langue

Comparaison de la **variance :** $p = 0.001$ *(PERMANOVA)*

Résultats - Taxons discriminants



Analyse linéaire discriminante d'effet de taille (*Linear discriminant analysis Effect Size, LefSE)*

→ Ces 9 genres sont ceux qui expliquent le plus la différence dans la composition des communautés entre les deux habitats.

→ **Mais <u>pourquoi</u>?**

Pour plus d'infos :
Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., & Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome biology*, *12*, 1-18.

Schéma récapitulatif

Région 1

x 18 individus

Région 2

Échantillonnage

Échantillon avec cellules bactériennes et autre

Extraction de l'ADN

Tous les gènes bactériens en faible quantité

PCR

Plusieurs copies de la région V5-V6 du gène ARNr 16S

Séquençage de l'amplicon 16S

Fichier FASTQ contenant la suite de nucléotide composant la région V5-V6 du gène ARNr 16S

Traitement bio-informatique

Matrice d'abondance des ASVs par échantillons

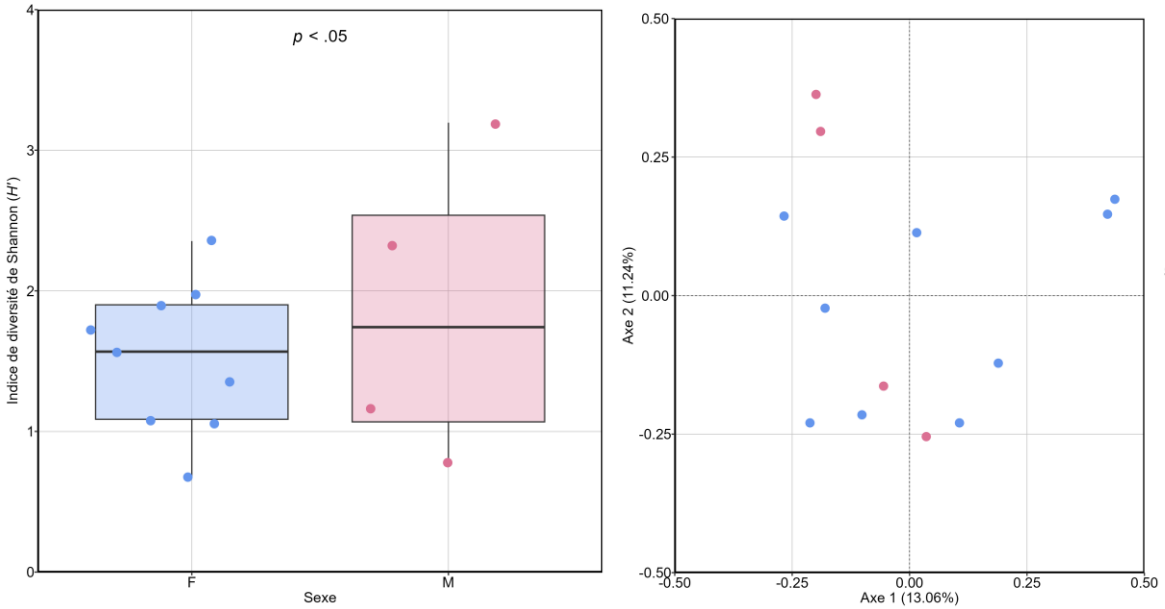
ASV	Phylum	Class	Ordre	Famille	Genre	Sample1	Sample2
ASV1	Firmicute	Bacilli	Saph	Staph	Saph	0	5
ASV2	Firmicute	Bacilli	Bacillales	Bacillalace	Bacillus	16	24
ASV3	Myxococo	Polyangia	Blfdi19	*Unclassified*	*Unclassified*	11	39
ASV4	Proteobac	Alphaproteo	R7C24	*Unclassified*	*Unclassified*	0	0
ASV5	Proteobac	Alphaproteo	R7C24	*Unclassified*	*Unclassified*	2	50

Exercice R

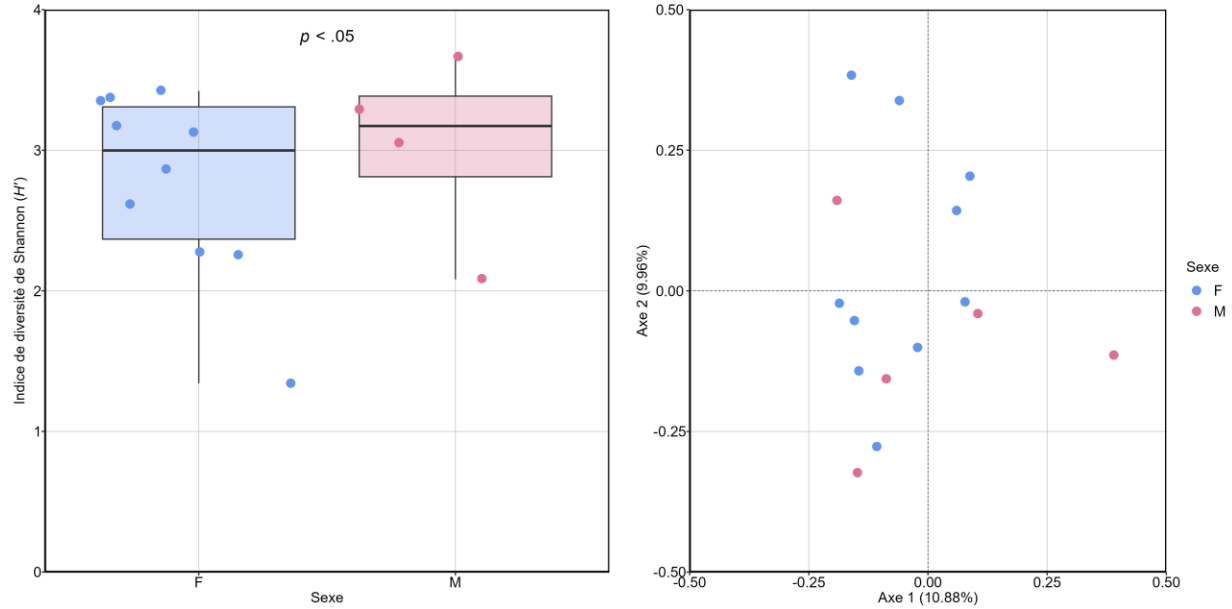
- Objectif :
  - Utiliser R Studio pour générer des graphiques à barres des genres microbiens les plus abondants sur votre langue et votre main
  - Utiliser Excel afin de récupérer la séquence du taxon le plus abondant sur votre langue et votre main pour l'analyse BLAST

Questions ?

Diversité et composition en fonction du sexe biologique à la naissance

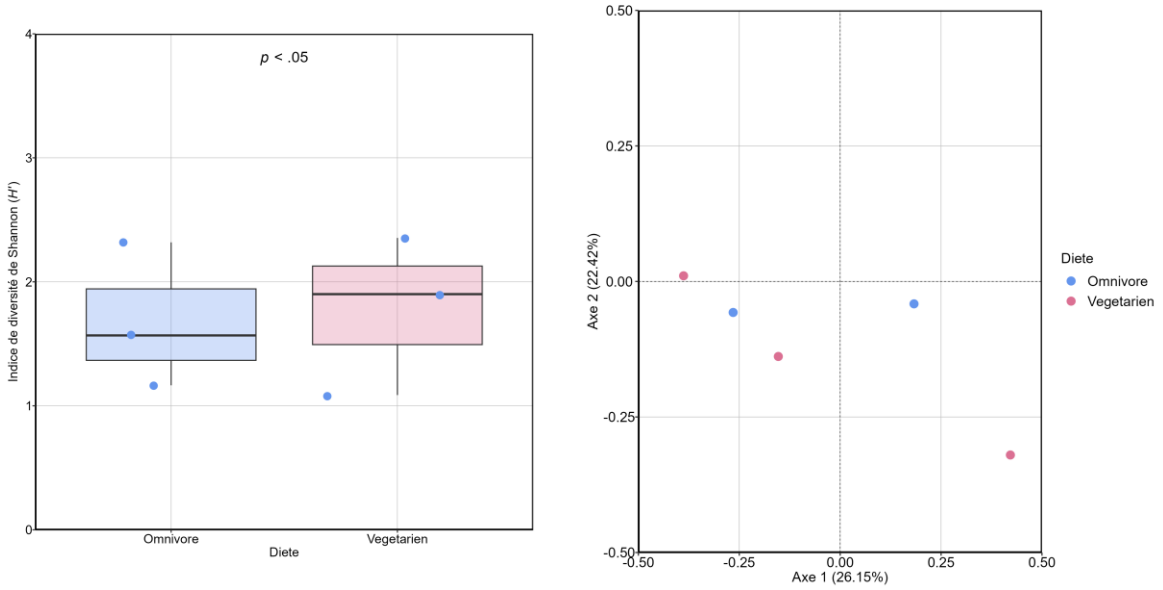


Langue

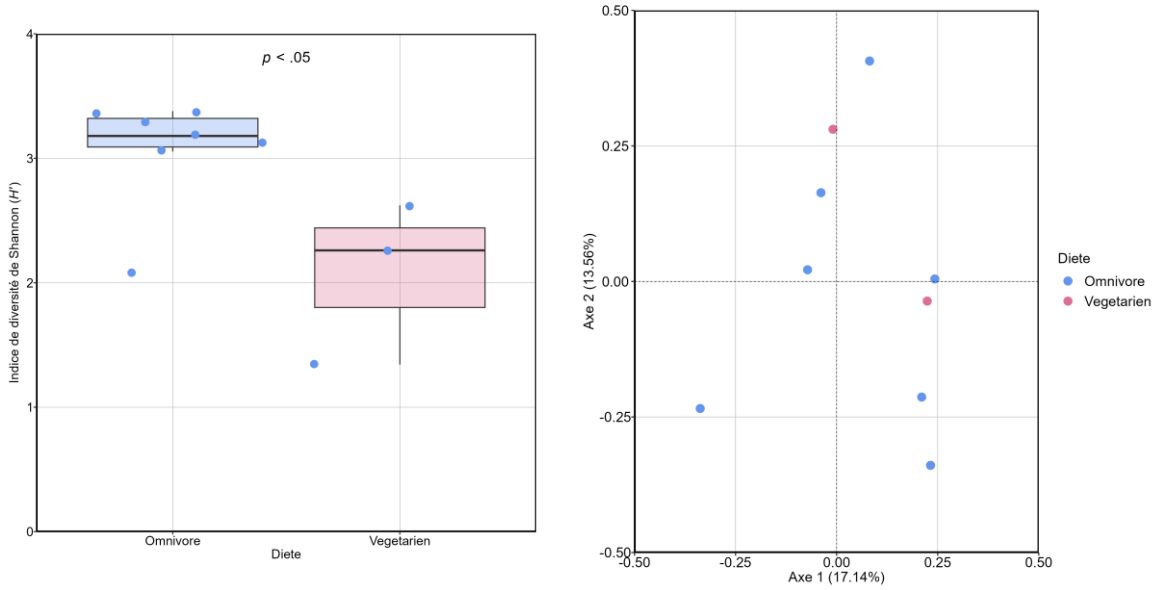


Main

Diversité et composition en fonction de la diète

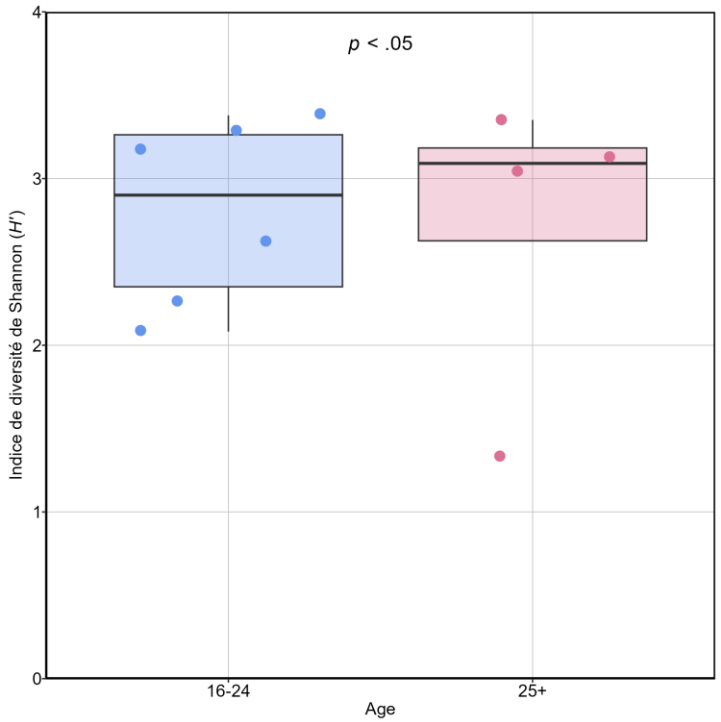


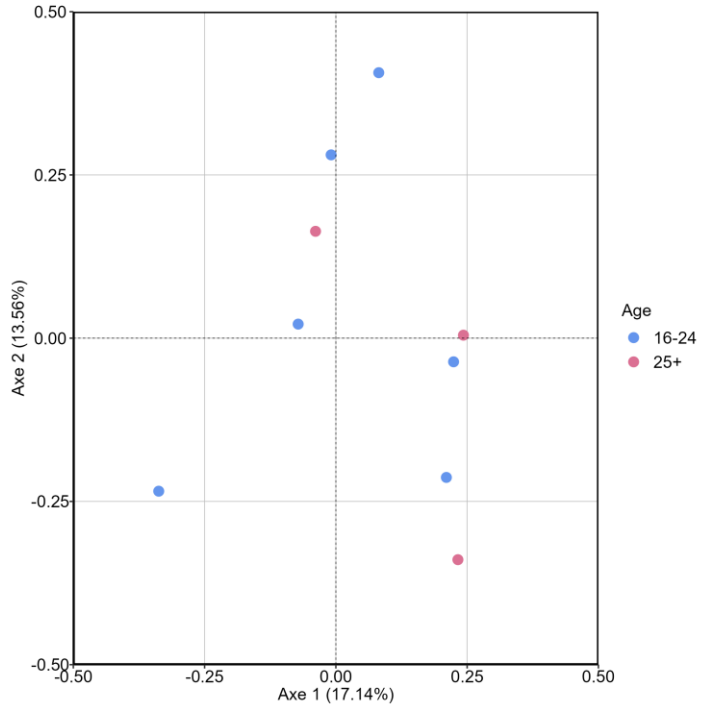
Langue



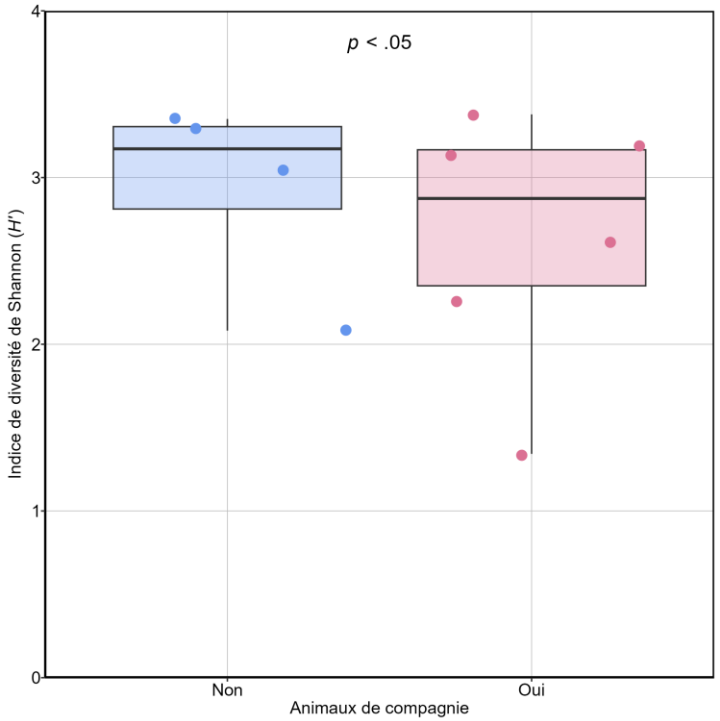
Main

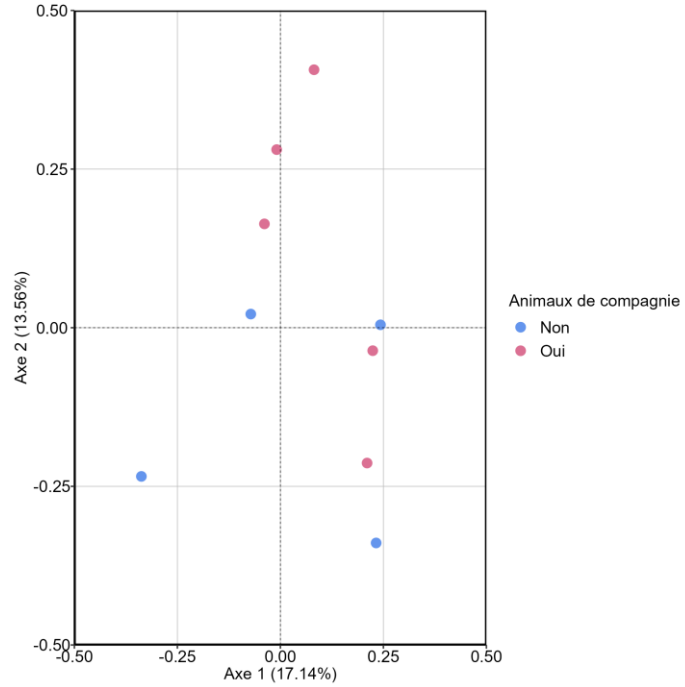
Diversité et composition en fonction de l’âge (main)





Diversité et composition en fonction de la présence des animaux de compagnie (main)





Questions avec n insuffisant

Lieu de residence

Method de transport

Consommation du café

Consommation de l’alcool

Lavage des mains

Main dominante

Brossage des dents

Activité*

Covid-19

Antibiotiques

Fumeur ou non

Presence des plantes*

Colocation