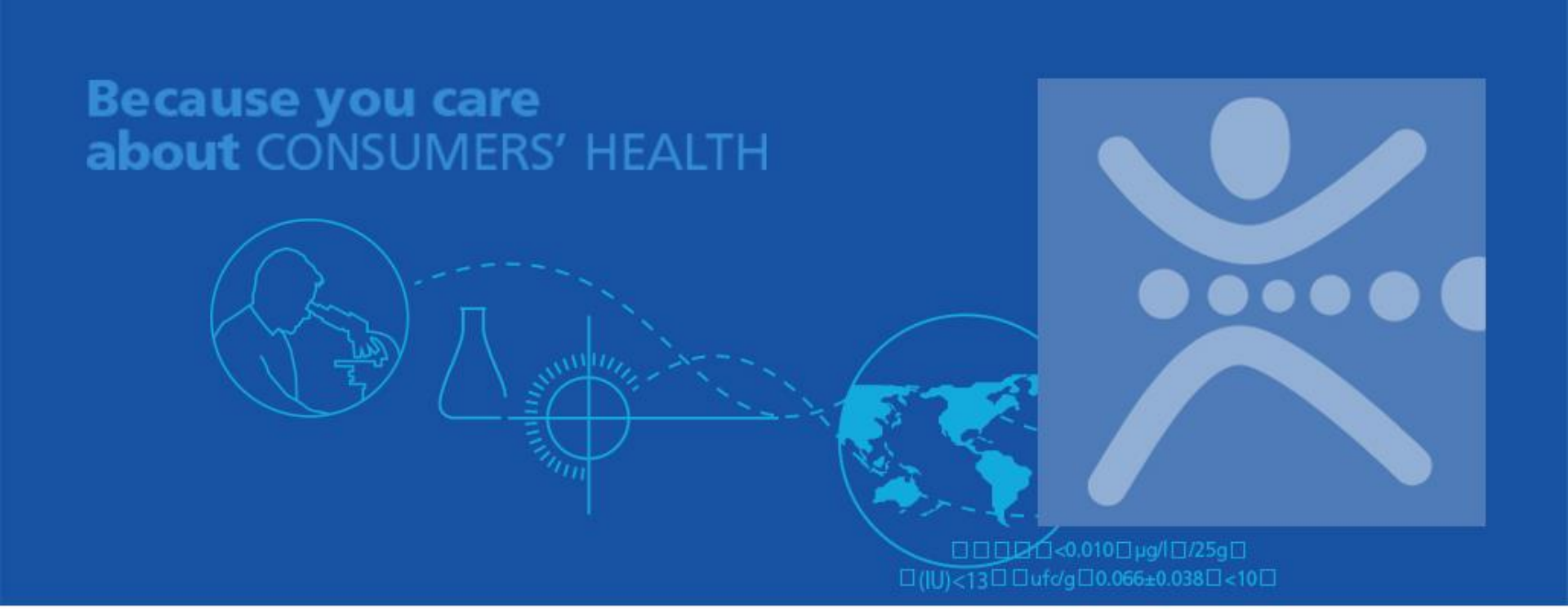


Méthodes de machine learning pour la recherche de signatures taxonomiques du microbiote intestinal

Application à des données issues de séquençage ciblé de régions variables des gènes de l'ARNr 16S

Webinaire Adebiotech - 15 Juin 2021

Florence Gillaizeau
PhD, Biostatistics Manager

ADEBIOTECH
THINK TANK ONE HEALTH

BIOFORTIS
a Mérieux NutriSciences Company

Introduction

Application : Méthodes

Application : Résultats

Conclusion

From Human to Data...
From Science to Market

BIOFORTIS INNOVATION SERVICES

CLINICAL EXPERTISE

MICROBIOME

CENTRAL LAB

DATA ANALYTICS

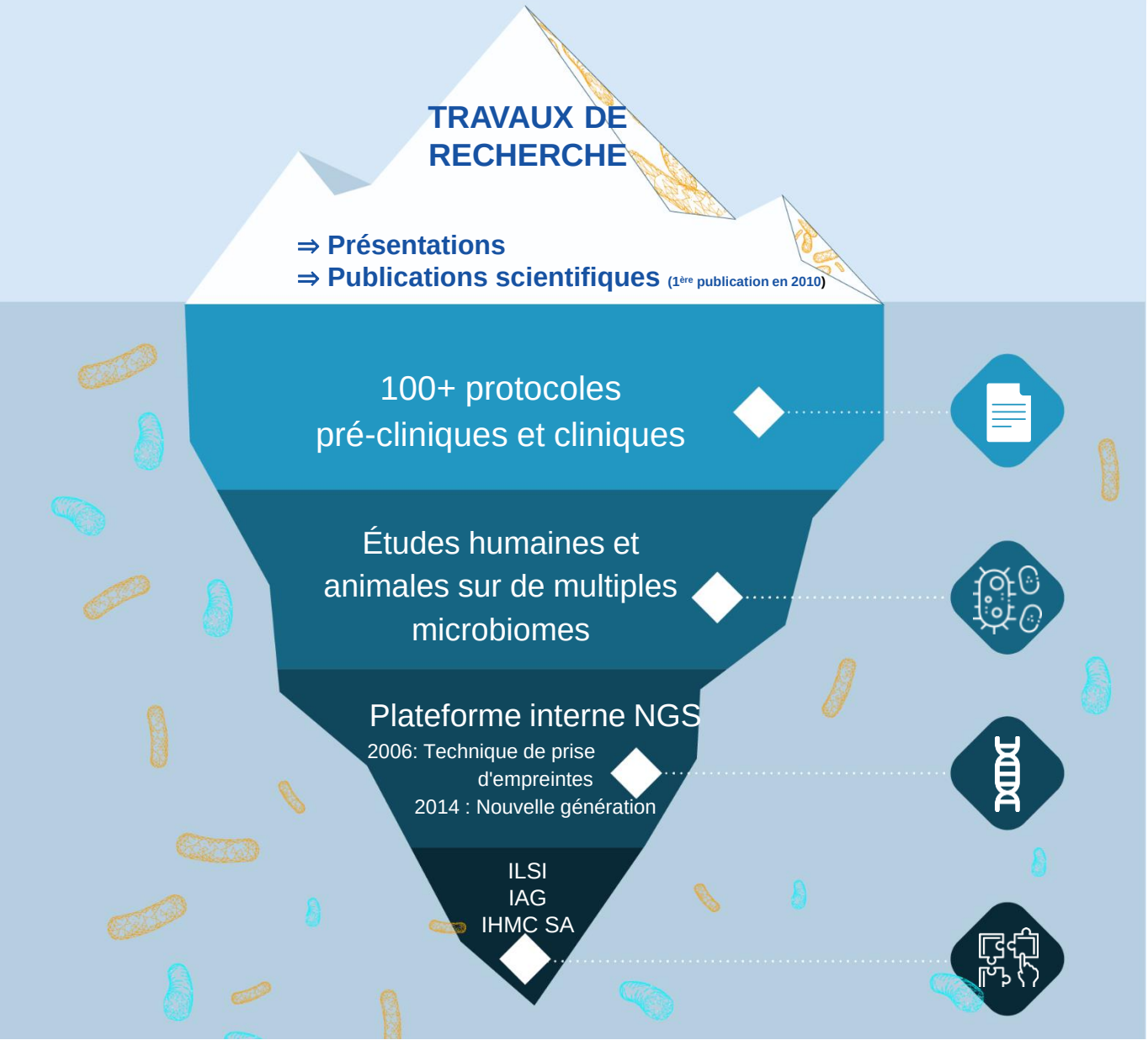
BIOFORTIS
a Mérieux NutriSciences Company

ICH
harmonisation for better health

QUALITY SYSTEM CERTIFICATION
DNV·GL
ISO 9001

TESTER LES EFFETS DES PRODUITS AVEC DES VOLONTAIRES, DES PATIENTS OU CONSOMMATEURS

Microbiome: Niveau d'expertise



Statistiques pour le microbiome

> Y-a-t-il un changement de profil du microbiome lié au traitement?

> Quelles espèces sont concernées par ce changement?

> Est-ce que le microbiome module l'effet d'un traitement?

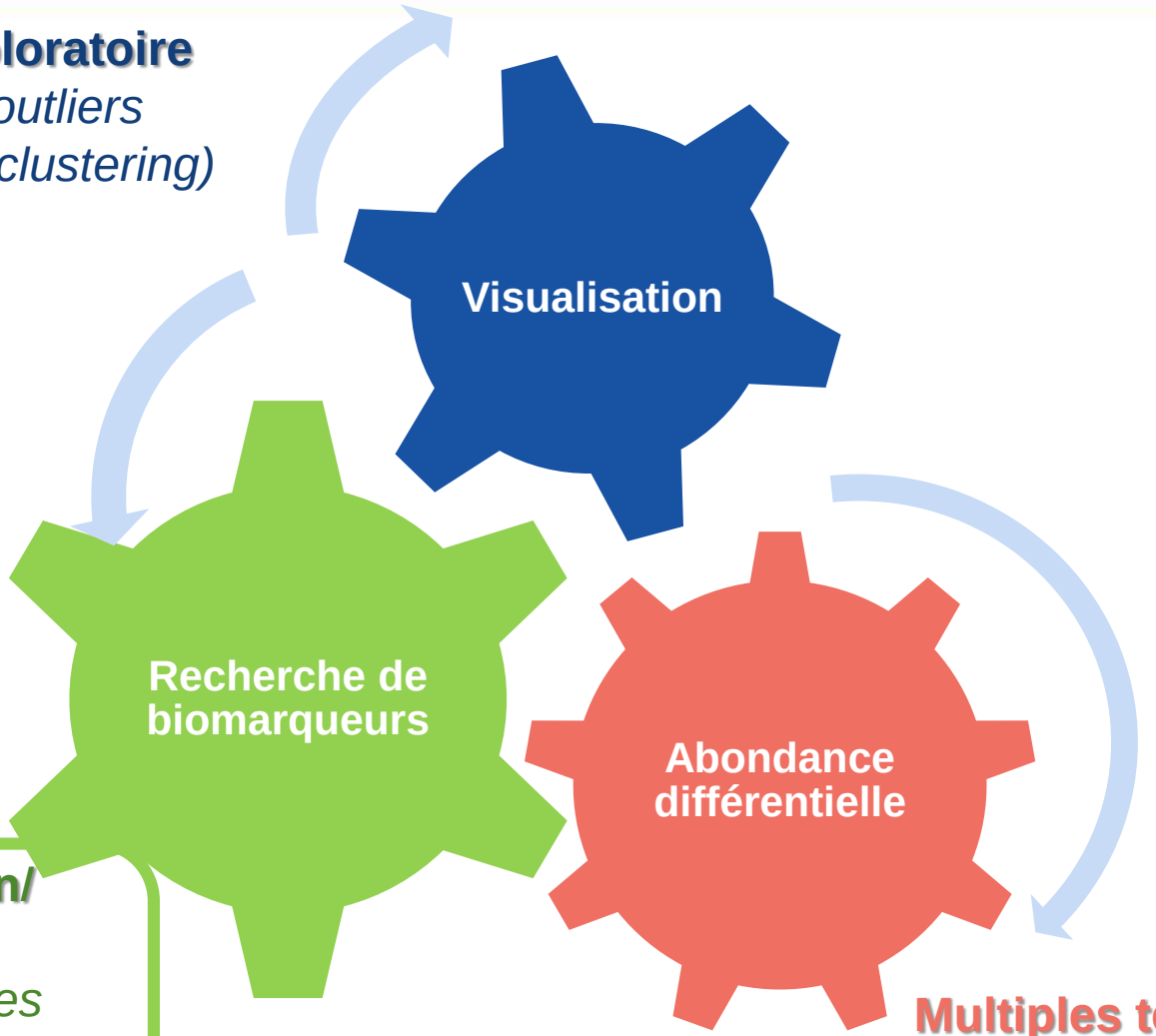
> Est-il possible de prédire une maladie ou un évènement à partir de signatures du microbiome?

> Peut-on établir une relation cause-effet?

"Data don't make any sense,
we will have to resort to statistics."

Statistiques pour le microbiome

Analyse exploratoire
- *Détection des outliers*
- *Classification (clustering)*



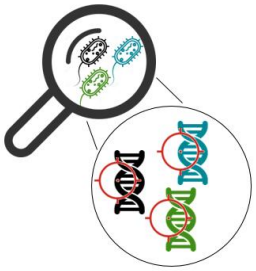
Test d'association/ Prédiction
- *Données longitudinales*
- *Ajustement de covariables*

Multiples tests d'hypothèses
- *Contrôle des faux positifs*
- *Puissance*

Deux approches principales de séquençage

Metabarcoding 16S

Séquençage Shotgun



Taxonomic Profile

	Samples			
	S_1	S_2	...	S_x
$Taxa_1$				
$Taxa_2$				
$Taxa_3$				
$Taxa_4$				
$Taxa_5$				
...				
$Taxa_y$				

Functional/Gene Profile

	Samples			
	S_1	S_2	...	S_x
$Gene_1$				
$Gene_2$				
$Gene_3$				
$Gene_4$				
$Gene_5$				
...				
$Gene_y$				

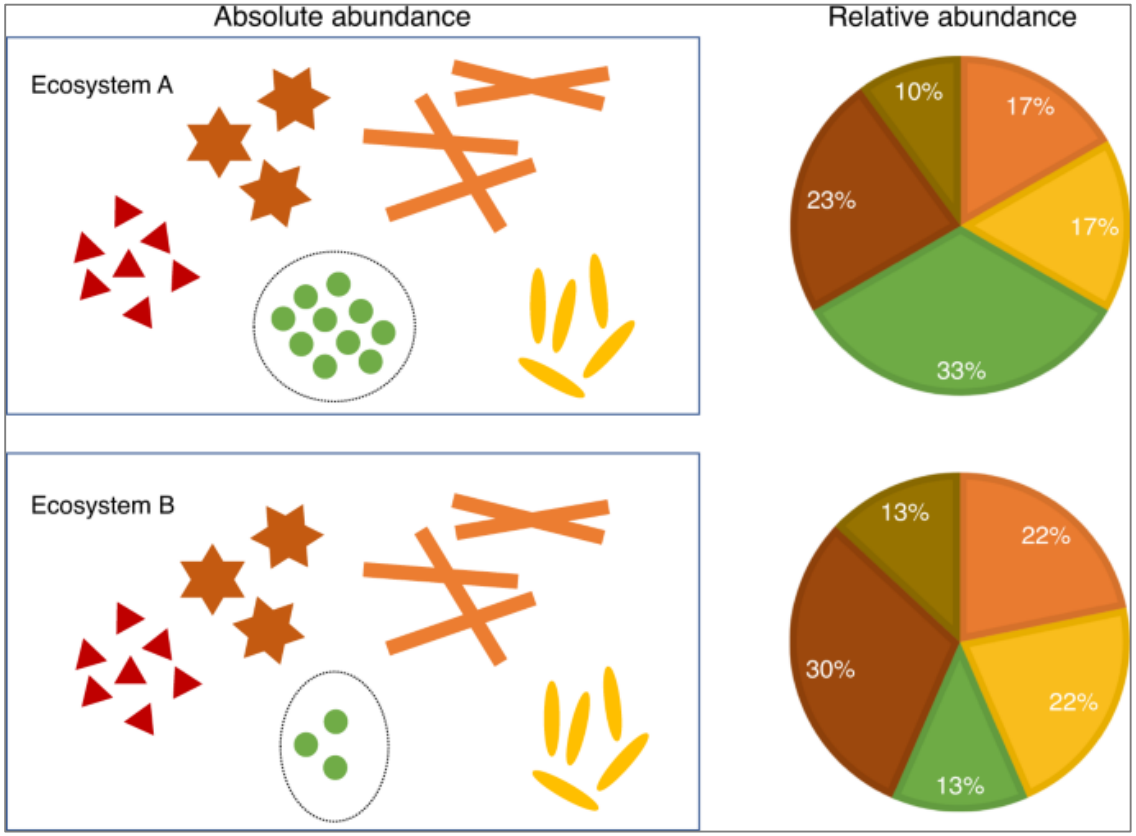
Qui est là ?

Que peuvent-ils faire?

Profondeur de séquençage

Critères d'évaluation du microbiome

- Composition du microbiome : abondances absolues/relatives



Lin, H., Peddada, S.D. Analysis of compositions of microbiomes with bias correction. *Nat Commun* **11,** 3514 (2020).

Critères d'évaluation du microbiome

- Composition du microbiome : abondances absolues/relatives
- Diversité du microbiome
  - au sein d'un échantillon (α-diversité)

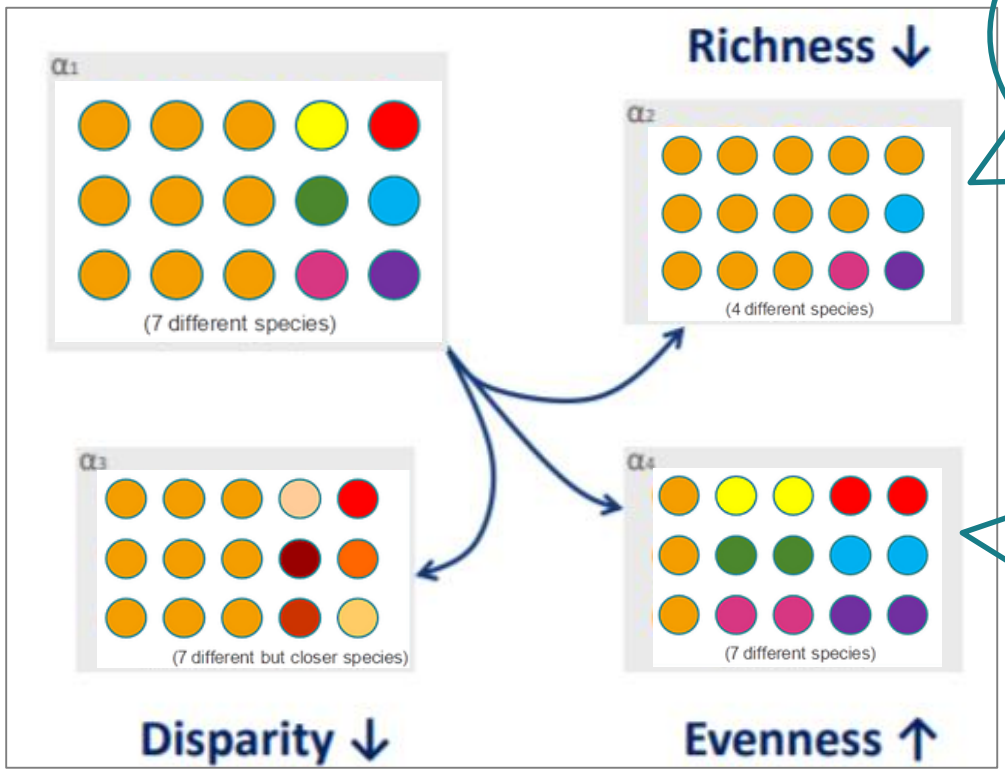


Illustration des indices de diversité (Thomas Carton).

Combien d'espèces différentes sont présentes dans mon échantillon?

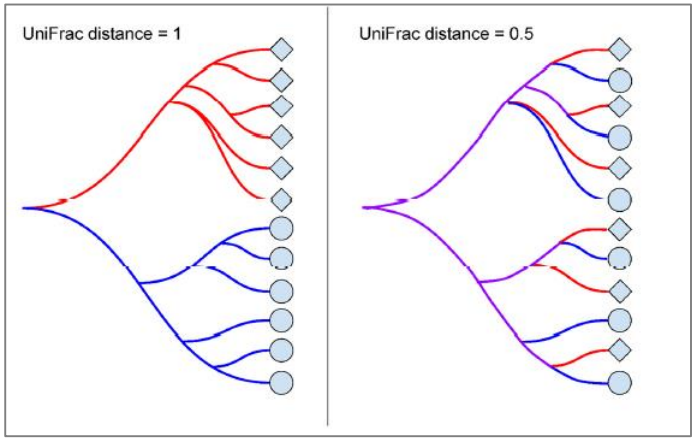
Y-a-t-il une répartition équitable des espèces ou certaines espèces sont-elles dominantes?

Critères d'évaluation du microbiome

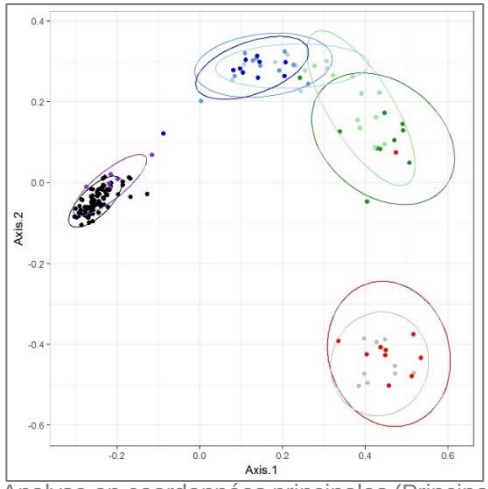
- Composition du microbiome : abondances absolues/relatives
- Diversité du microbiome
  - au sein d'un échantillon (α-diversité)
  - entre les échantillons (β-diversité)

Quelle différence de composition du microbiome entre groupes (différents traitements, temps, …)?

- Distance de Jaccard : basée sur la **présence/absence** de taxa
- Distance de Bray-Curtis : basée sur les **abondances** de taxa
- Distance UNIFRAC non pondérée : basée sur la **présence/absence** de taxa et la **phylogénie**
- Distance UNIFRAC pondérée: basée sur les **abondances** de taxa et la phylogénie

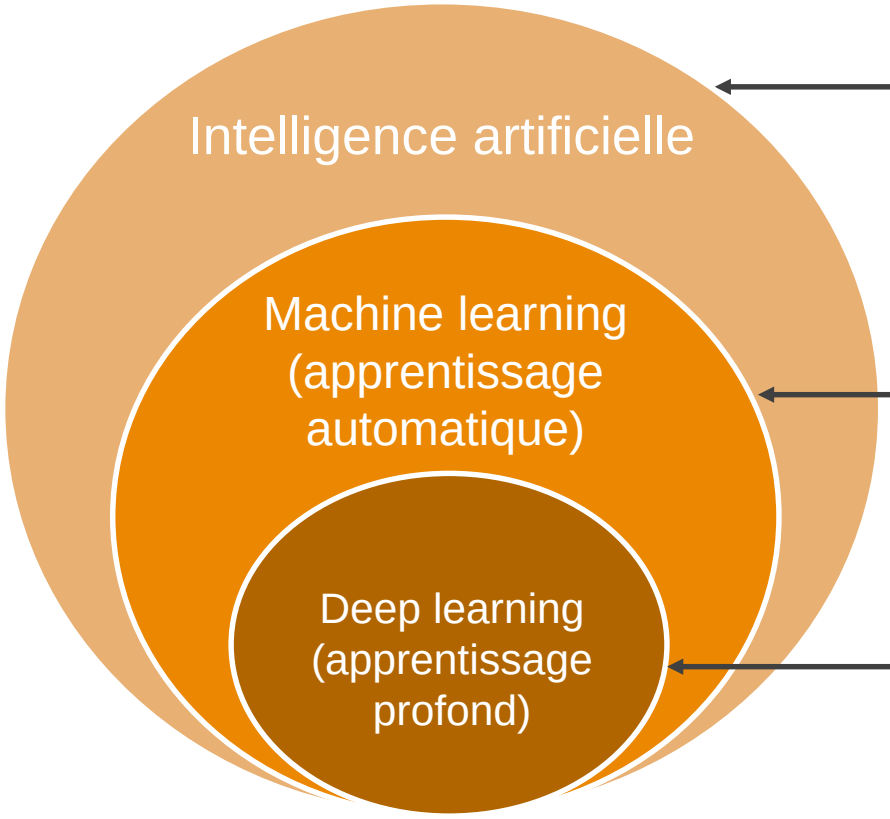


Wong RG, Wu JR, Gloor GB (2016) Expanding the UniFrac Toolbox. PLoS ONE 11(9):e0161196.



Analyse en coordonnées principales (Principal Coordinates Analysis (PCoA)) pour explorer et visualiser les similarités entre échantillons.

Place du machine learning (ML)



Techniques permettant à des machines de mimer le comportement humain

Utilisation d'algorithmes stats pour apprendre à partir des données puis en extraire de l'information / réaliser des prédictions

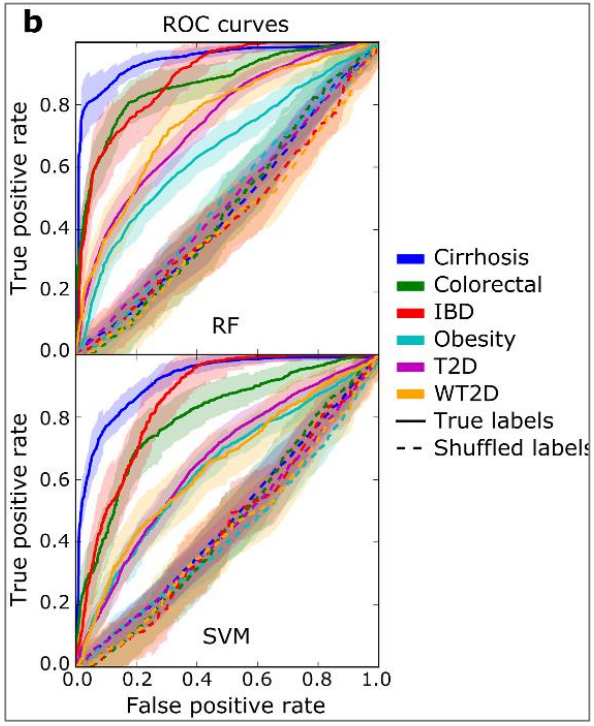
Basé sur réseaux de neurones multi-couches

Machine learning (ML) en pratique

- Utilisation : assister les professionnels de santé pour le diagnostic ou le pronostic
  - ➔ vers une médecine des 5P : Préventive, Prédictive, Participative, Personnalisée et Pertinente

- Algorithmes commun de ML :
  - **random forest** (forêt aléatoire),
  - SVM (support vector machine),
  - elastic net,
  - **LASSO** (Least Absolute Shrinkage and Selection Operator),
  - …

- Applications liées au microbiome :
  - combiner des informations du microbiote pour prédire l'état de santé actuel ou futur (score ou statut répondeur/non répondeur) d'un individu



Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights. *PLoS Comput. Biol.* 12:e1004977.

Introduction

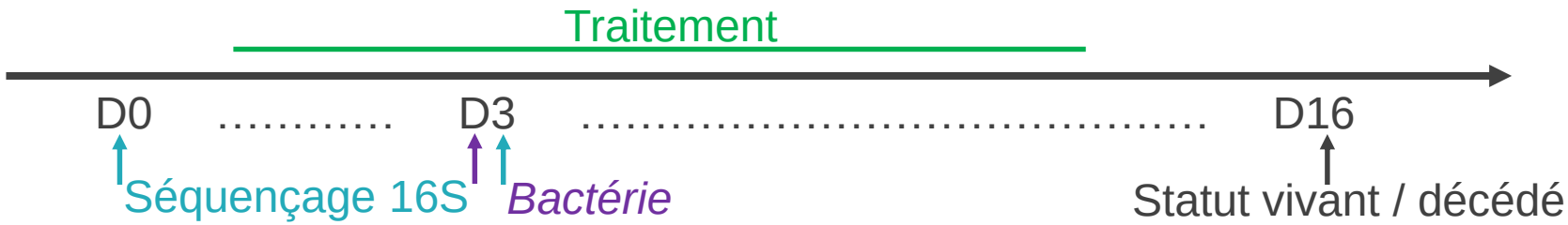
Application : Méthodes

Application : Résultats

Conclusion

Application à une étude pré-clinique

- Modèle murin (n=60 animaux)
  - Innoculés avec bactérie pathogène causant des diarrhées sévères
  - Soumis à un traitement



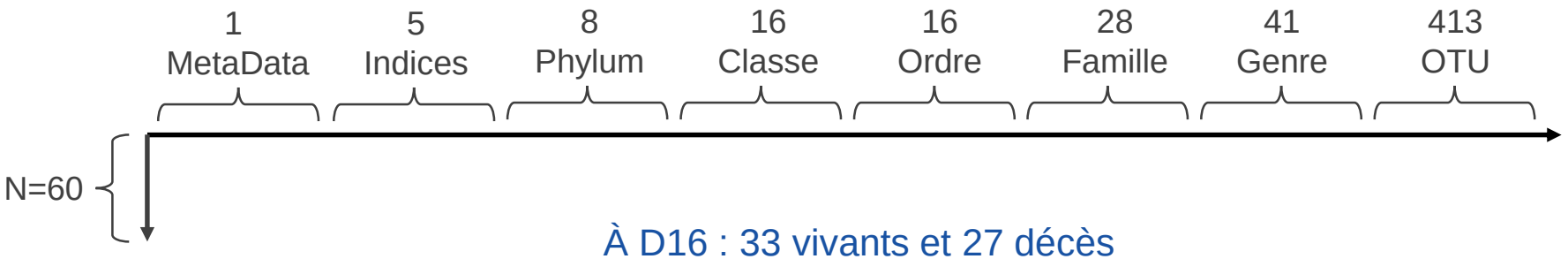
- **Objectif** : Identifier des signatures taxonomiques bactériennes du microbiote intestinal (taxa et/ou indices d'alpha diversité) issus de données de séquençage ciblé de régions variables du gène de l'ARNr 16S qui prédisent le décès des animaux.

Données

- Données 16S brutes : approche metabarcoding ciblant la région V3-V4 du gène codant pour l'ARNr 16S (technologie Illumina)

  - Normalisation (abondances relatives log-transformées)
  - Variation D3-D0

- Données



À D16 : 33 vivants et 27 décès

Analyses

Données étude → Statistiques descriptives

x 6 niveaux

- Phylum
- **Classe**
- Ordre
- Famille
- Genre
- OTU

x 2 sets de variables

- **Taxa**
- Taxa + Indices (Shannon index, Number of observed, OTU, Inverse Simpson index, Unweighted UniFrac distance, Bray-Curtis dissimilarity)

12 jeux de données

x 2 méthodes

- LASSO
- Random Forest

24 modèles → Sélection de variables et importance pour le statut à D16
Qualité de prédiction du modèle

Machine learning
- LASSO (Least Absolute Shrinkage and Selection Operator)

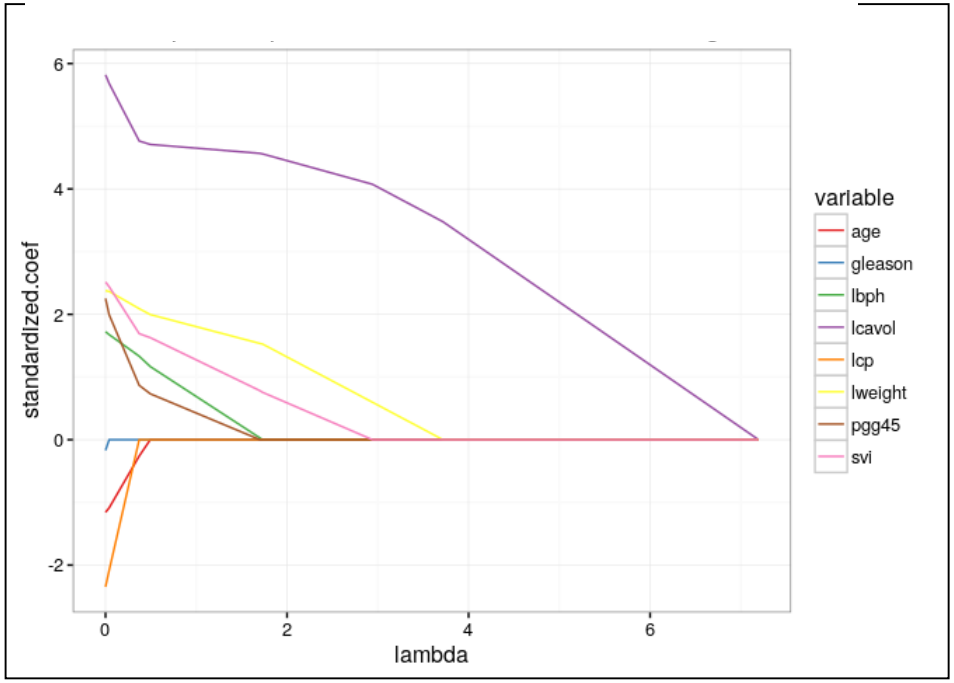
Comment prédire le risque de décès à partir d'une **combinaison linéaire de prédicteurs** du microbiote intestinal?

- Modèle : Régression logistique

$$\log\left(\frac{Proba(décédé\ à\ J16)}{Proba(vivant\ à\ J16)}\right) = \beta_0 + \beta_1 Taxon_1 + \beta_2 Taxon_2 + \ldots + \beta_k Taxon_k$$

- Sélection des variables: pénalisation LASSO *(Tibshirani, 1996)*

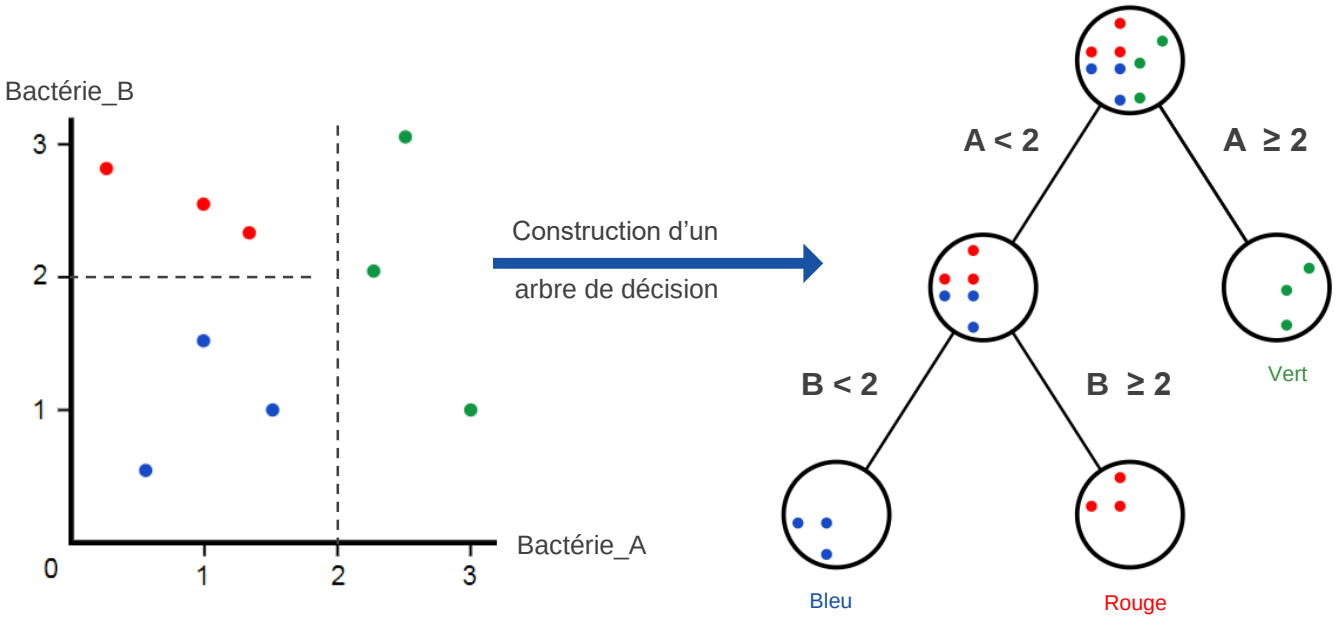
- Réduit le nombre de taxa du modèle en introduisant un paramètre de pénalisation λ
- Le paramètre λ
  - fait tendre vers 0 les coefficients des taxa les moins prédicteurs
  - est calculé par un algorithme d'optimisation minimisant l'erreur de prédiction de décès



Machine learning
- Random Forest

Comment prédire le risque de décès à partir d'une **combinaison d'arbres de décision** construits à partir de caractéristiques du microbiote intestinal?

- Algorithme : Random forest (forêt aléatoire) *(Breiman, 2001)*
  - Tirage aléatoire des individus et d'un sous-ensemble de variables à chaque séparation (noeud).
  - Collection d'arbres de décision.

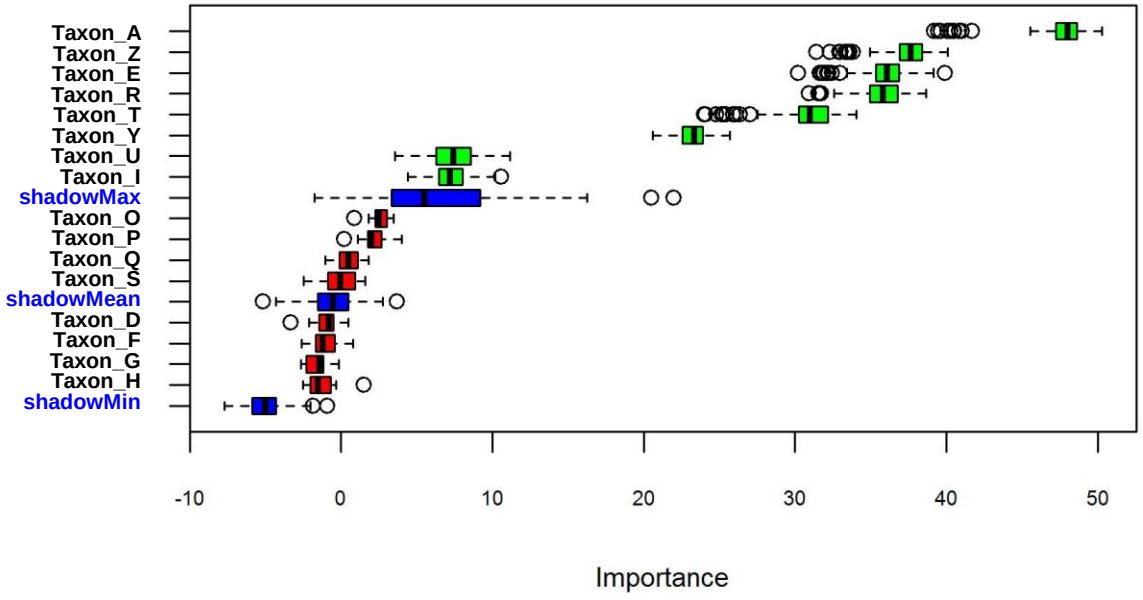


Machine learning
- Random Forest

Comment prédire le risque de décès à partir d'une **combinaison d'arbres de décision** construits à partir de caractéristiques du microbiote intestinal?

- Algorithme : Random forest (forêt aléatoire) *(Breiman, 2001)*
  - Tirage aléatoire des individus et d'un sous-ensemble de variables à chaque séparation (noeud).
  - Collection d'arbres de décision.

- Sélection des variables : algorithme de Boruta *(Degenhardt et al., 2017)*
  - Comparaison de l'importance des variables avec des variables aléatoires fictives en utilisant des tests statistiques et plusieurs iterations de forêts aléatoires.



Introduction

Application : Méthodes

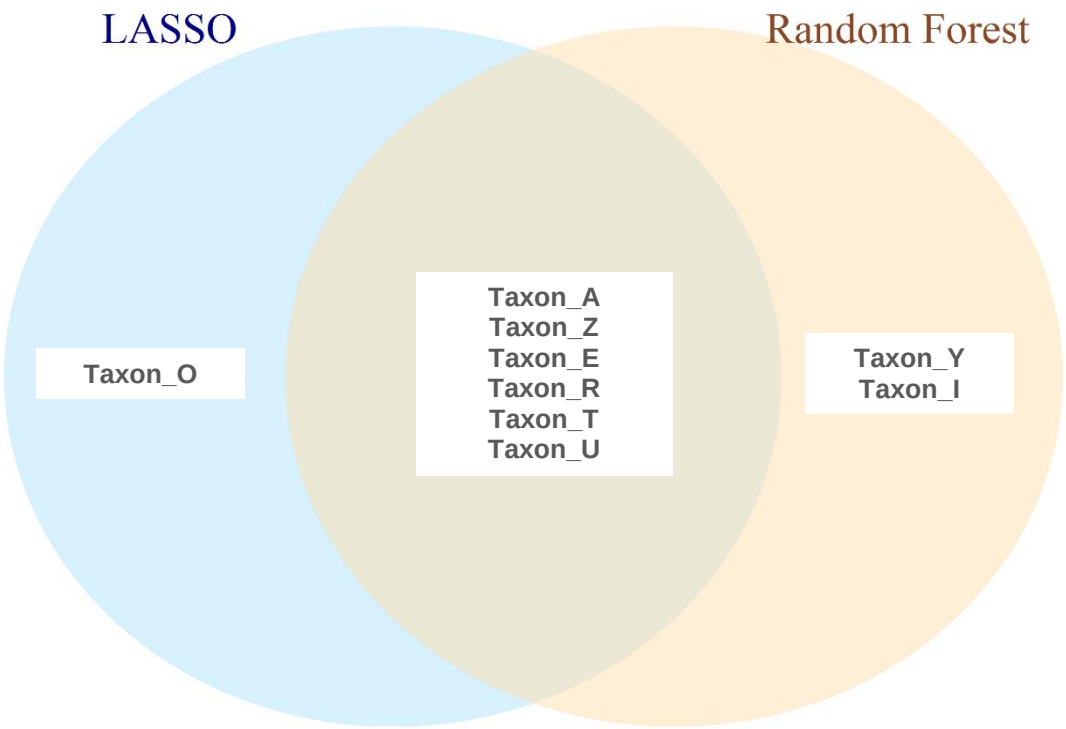
Application : Résultats

Conclusion

Résultats
- Exemple : Classe

- Taxa sélectionnés selon les méthodes

Méthode	Nombre de taxa sélectionnés	Taxa
LASSO	7	**Taxon_A**, **Taxon_Z, Taxon_E, Taxon_R, Taxon_T, Taxon_U,** Taxon_O,
Random Forest	8	**Taxon_A**, **Taxon_Z, Taxon_E, Taxon_R, Taxon_T,** Taxon_Y, **Taxon_U,** Taxon_I



Rappel taxonomie

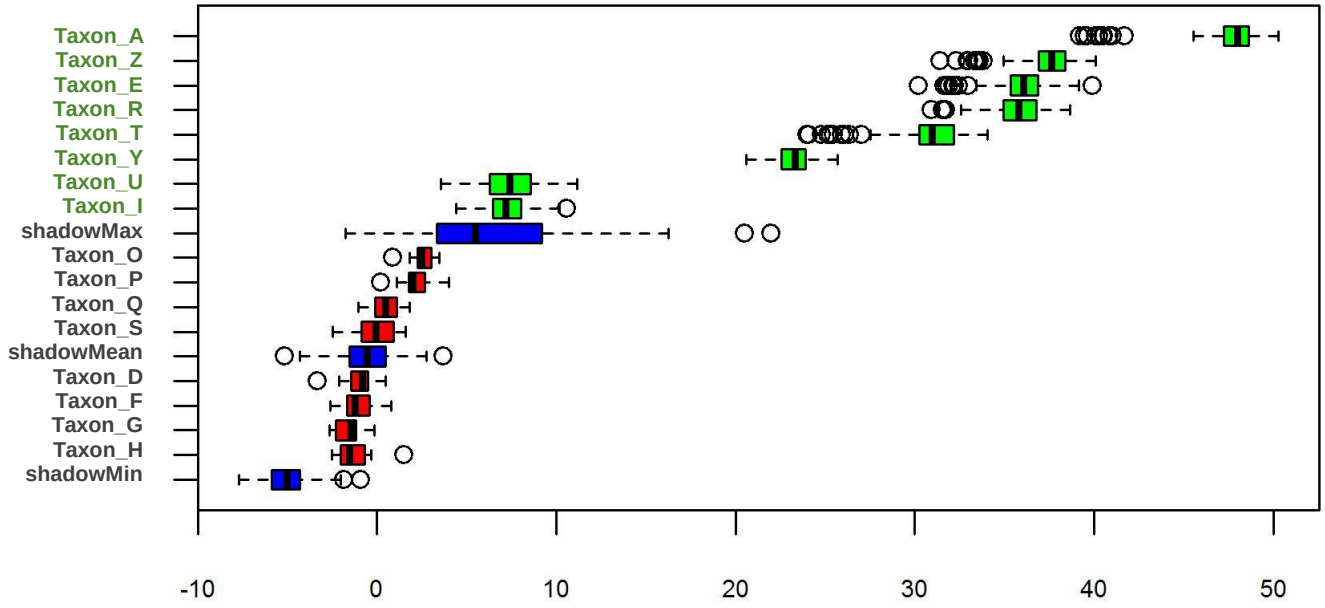
Phylum (8)
Classe (16)
Ordre (16)
Famille (28)
Genre (41)
OTU (413)

Résultats
- Exemple : Classe

- Informations sur les taxa sélectionnés
  - Coefficients LASSO

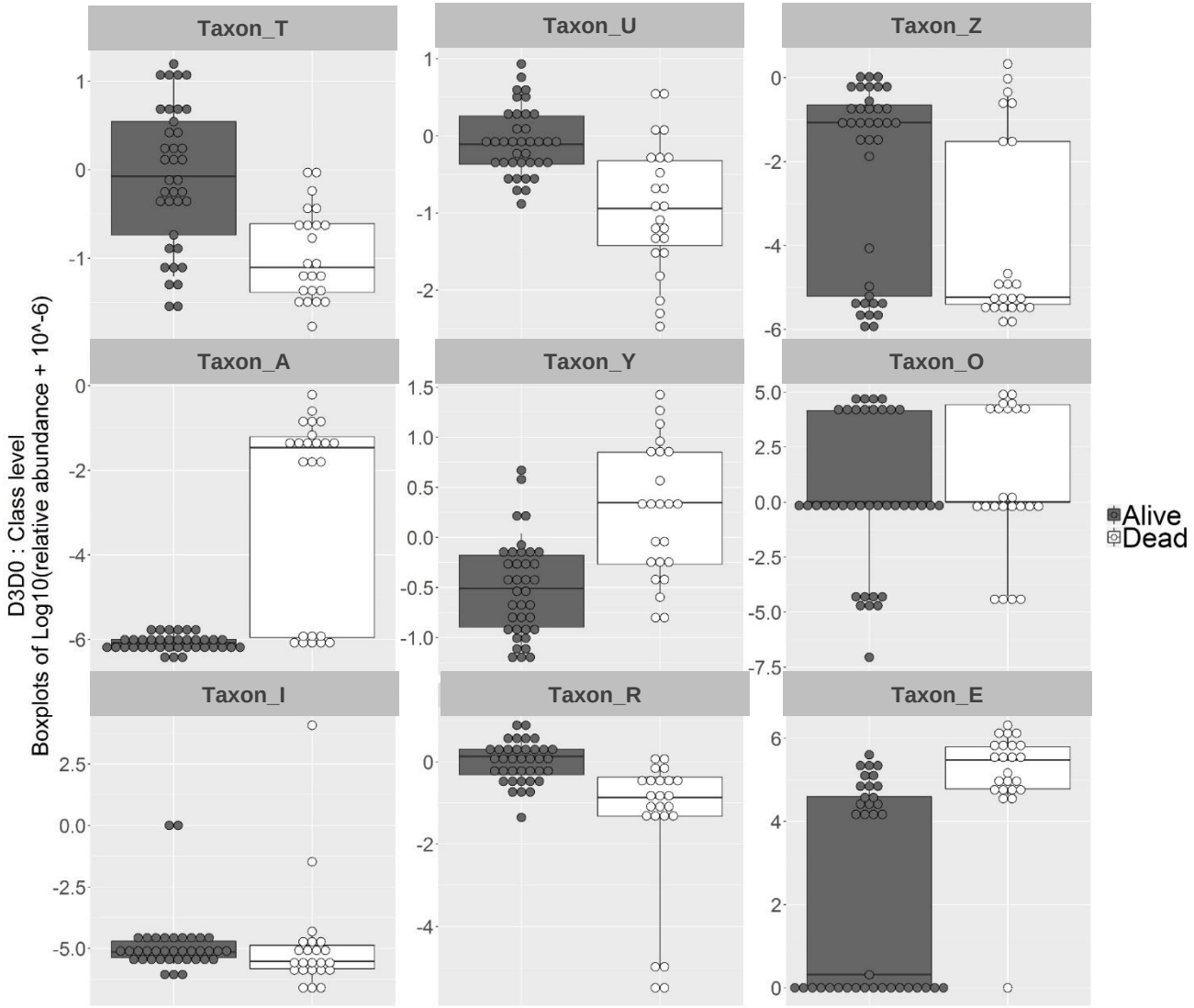
Taxa	Coefficient
(Intercept)	-0,070
Taxon_T	-1,502
Taxon_Z	-1,387
Taxon_R	-0,591
Taxon_U	-0,118
Taxon_O	0,055
Taxon_E	0,203
Taxon_A	0,720

  - Importance Random Forest



Résultats
- Exemple : Classe

- Association des variables au statut à D16
  - Boxplot des groupes d'animaux vivants / décédés

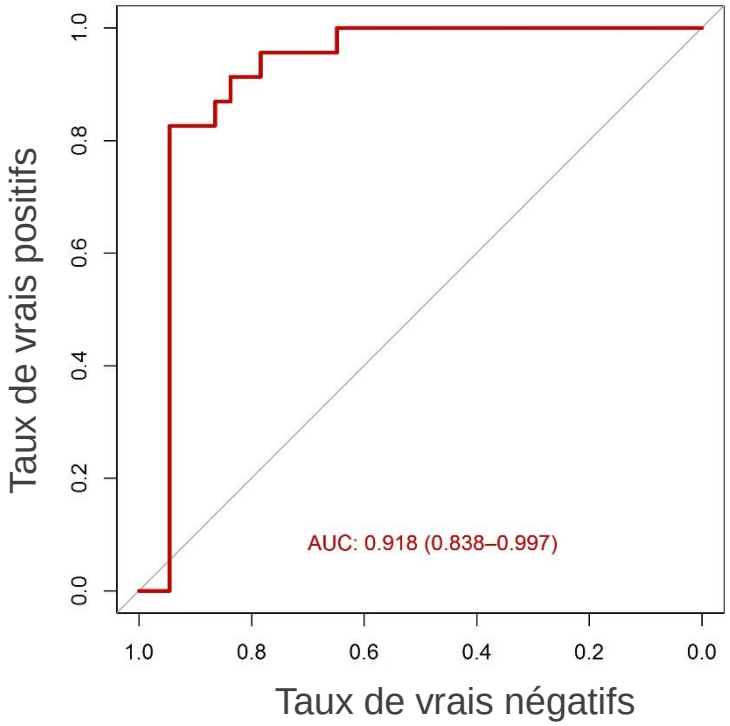


Résultats
- Exemple : Classe

- Évaluation de la qualité de prédiction du modèle

Méthode	Nb taxa	Taxa	AUC	95 % CI AUC
LASSO	7	**Taxon_A**, **Taxon_Z, Taxon_E, Taxon_R, Taxon_T, Taxon_U,** Taxon_O,	0.918	[0.838 ; 0.997]
Random Forest	8	**Taxon_A**, **Taxon_Z, Taxon_E, Taxon_R, Taxon_T,** Taxon_Y, **Taxon_U,** Taxon_I	0.881	[0.794 ; 0.967]

- Courbe ROC LASSO



Résultats
- Résumé des différents niveaux taxonomiques

- Nombre de variables sélectionnées par chaque modèle et qualité de prédiction :

Variables candidates	LASSO		Random Forest	
	Nb var/Nb var total	AUC	Nb var/Nb var total	AUC
Phylum	5/8	0.886	7/8	0.775
Phylum + Indices	10/13	0.972	10/13	0.797
Classe	7/16	0.918	8/16	0.881
Classe + Indices	13/21	0.976	11/21	0.810
Ordre	7/16	0.918	8/16	0.859
Ordre + Indices	13/21	0.976	11/21	0.824
Famille	9/28	0.989	13/28	0.856
Famille + Indices	11/33	1.000	17/33	0.821
Genre	11/41	0.979	16/41	0.864
Genre + Indices	9/46	0.944	21/46	0.851
OTU	18/413	1.000	97/413	0.965
OTU + Indices	20/418	1.000	94/418	0.965

- Bonne qualité de prediction (>0.8) avec les deux méthodes.
- La méthode LASSO a tendance à sélectionner moins de variables (mais ne convergeait pas pour tous les modèles).

Introduction

Application : Méthodes

Application : Résultats

Conclusion

Conclusion

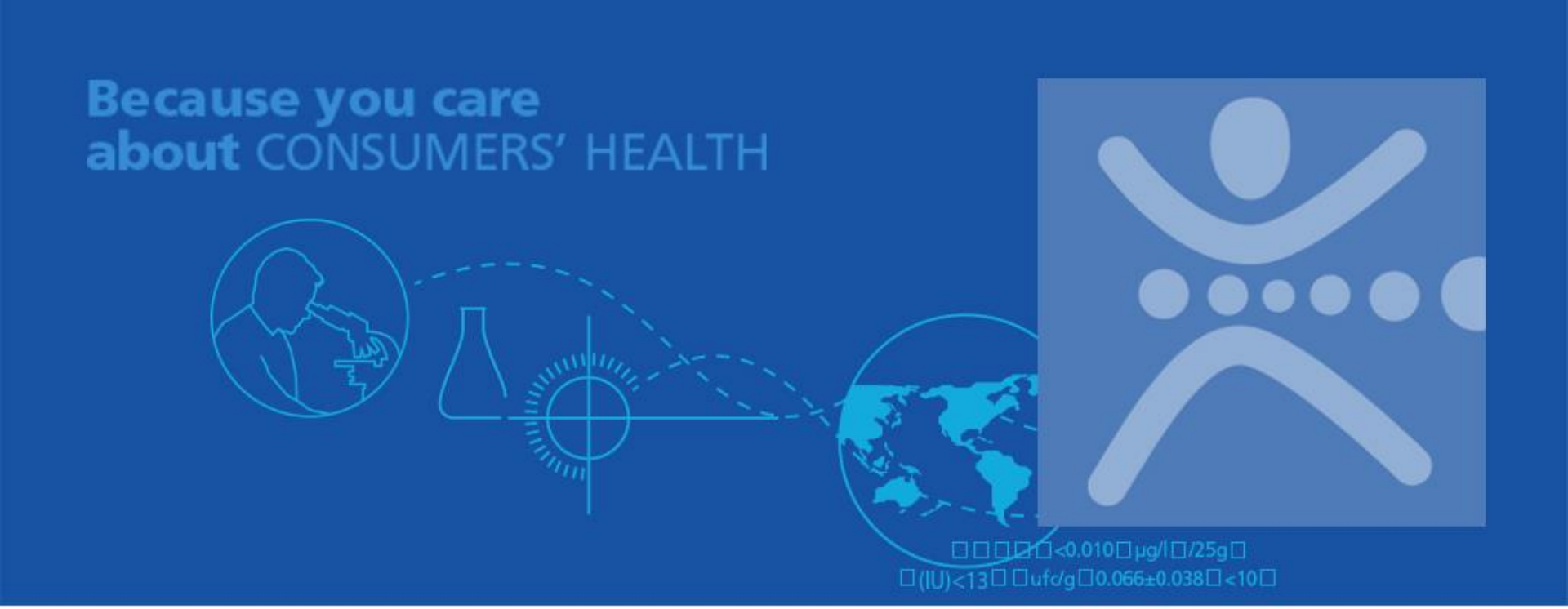
- Méthodes de machine learning
  - Algorithmes d'apprentissage permettant de classer ou prédire.
  - Nombre de variables explicatives >> nombre d'observations.

  ➔ Adaptées au contexte du microbiome : Nombre de taxa, indices >> nombre d'échantillons, règle de décision complexe.

- Exemple discuté
  - Peut être étendu à d'autres types de données de séquençage (Shotgun).
  - La signature peut également intégrer d'autres types de variables (biologiques, génétiques,…).
  - La robustesse des algorithmes peut être limitée par la taille d'échantillon et/ou la dépendance des variables ➔ estimer la qualité de prédiction du modèle (validation interne, validation externe).

Challenges

- L'application des méthodes de machine learning aux données du microbiome est récente ➔ restent de nombreux challenges
  - ➔ Nombreuses sources de variations : outils bioinformatiques et pipelines, normalisation des données, variabilité entre individus, algorithmes, …
  - ➔ Design des études et corrélations entre variables (données longitudinales, plusieurs sites de prélèvements,…).
  - ➔ En 2019, création du réseau européen COST (European Cooperation in Science and Technology) Action "ML4Microbiome" (Machine Learning for Microbiome). *(Moreno-Indias et al., 2021)*



Merci !

biofortis.merieuxnutrisciences.com

BIOFORTIS
a Mérieux NutriSciences Company