Software engineering and genericity of a transgenerational hologonomic simulation framework, followed by an in silico analysis of thermal effects on the bovine holobiont

Youna Maillié, Mahendra Mariadassou, Andrea Rau, Ingrid David, Solène Pety

Ingénierie logicielle et généricité d'un cadre de simulation hologénomique transgénérationnelle, suivi d'une analyse des effets thermiques *in silico* sur l'holobionte bovin

Youna Maillié

Master AMIIB - Université Paris-Saclay

Encadrants :
Solène Pety (INRAE, GABI et MaIAGE, équipe StatinfOmics)
Andrea Rau (INRAE, GABI, équipe GiBBS)
Mahendra Mariadassou (INRAE, MaIAGE, équipe StatinfOmics)

Tuteur académique :
Philippe Rinaudo

Référente universitaire :
Sarah Cohen-Boulakia (LISN, Université Paris-Saclay)

30 août 2025

Ce rapport est le fruit des activités que j'ai mené au cours de mon stage de fin d'études, du Master Analyse Modélisation et Ingénierie de l'Information Biologique et Médicale de l'Université Paris-Saclay.

Je tiens ici à remercier l'ensemble des personnes ayant rendu possible la concrétisation de ce travail, aboutissement de mon parcours universitaire.

Tout d'abord, je souhaite particulièrement remercier mes trois encadrants pour leur accompagnement tout au long de mes travaux. Mahendra Mariadassou pour m'avoir apporté un savoir et une réflexion scientifique des plus avisées. Andrea Rau pour sa méthodologie et ses précieuses recommandations en analyse statistique. Solène Pety, pour avoir fait de moi une stagiaire intégrée et riche d'un soutien solide face aux aléas des données ; merci de m'avoir accordé ta confiance au regard de ce que RITHMS représente pour toi. Merci à vous trois pour les relectures de ce mémoire et la qualité de vos conseils.

Je remercie l'ensemble des personnes que j'ai côtoyé au cours de mon aventure à MaIAGE, lors de discussions scientifiques ou plus informelles. Merci à Hajar, Xinzhi et Sylvain avec qui j'ai partagé mon espace de travail et pu échanger sur des sujets très variés. Je n'oublie pas les moments riches en émotions au premier étage et je vous en remercie pour ça : Oriane, Manon, Naïa, Hanin et Emilie-Jeanne.

Merci à toi Emma qui m'a permis de passer deux années incroyables dans le Master BIBS, et pour tous ces projets en binôme que l'on a réussi à construire !

Enfin, je souhaite remercier très sincèrement mes parents et les membres de ma famille pour leur soutien, même à distance, dans ces cinq années d'étude. Face aux moments de doutes, aux problèmes du quotidiens, et pour vos visites en région parisienne, merci d'avoir été présents et de m'avoir accompagnée jusqu'à la réussite de mon parcours universitaire.

Table des matières

Liste des Figures

Introduction

1.1 Environnement de travail

1.1.1 INRAE Jouy-en-Josas-Antony

J'ai intégré en mars 2025 l'Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE) sur le site de Jouy-en-Josas-Antony dans les Yvelines (78). Pour une durée de 6 mois, j'ai rejoint l'unité de Mathématiques et Informatique Appliquées du Génome à l'Environnement (MaIAGE), au sein de l'équipe StatInfOmics. Mon stage s'est déroulé sous la direction de Solène Pety avec un co-encadrement de Mahendra Mariadassou et Andrea Rau.

MaIAGE est une unité de recherche pluridisciplinaire composée de 4 équipes de recherche ainsi qu'une plateforme de bioinformatique : Dynenvie (modélisation dynamique et statistique pour les écosystèmes, l'épidémiologie et l'agronomie), Bibliome (acquisition et formalisation de connaissances à partir de textes), BioSys (biologie des systèmes), StatInfOmics (bioinformatique et statistiques des données omiques) et Migale (plateforme de bioinformatique). Elle dépend des départements Mathématiques et Numériques (MathNum, le département pilote) et Microbiologie et Chaîne Alimentaire (MICA). Elle regroupe des mathématiciens, des statisticiens et des bioinformaticiens contribuant au développement de méthodes mathématiques et informatiques de portée générique ou motivées par des problèmes biologiques précis. Au 1er Janvier 2025, l'unité est composée de 70 membres répartis entre les 5 équipes et accueille chaque année des stagiaires du L3 au M2 et des doctorants. Elle s'implique aussi dans la mise à disposition de base de données et de logiciels permettant aux biologistes d'utiliser des outils dans de bonnes conditions ou d'exploiter automatiquement la littérature scientifique. L'inférence statistique et la modélisation dynamique sont des compétences fortes de l'unité, auxquelles s'ajoutent la bioinformatique, l'automatique et l'algorithmique.

Au sein de l'unité MaIAGE, l'équipe StatInfOmics, animée par Mahendra Mariadassou, aborde des questions biologiques qui concernent principalement l'annotation structurale et fonctionnelle des génomes, les régulations géniques, la dynamique évolutive des génomes et la caractérisation d'écosystèmes microbiens en termes de diversité et de fonctions présentes ; une cible commune étant la relation entre génotype et phénotype. Par ailleurs, une part de plus en plus importante de son activité est liée à l'intégration de données omiques hétérogènes pour en extraire de l'information pertinente et également prédire des processus biologiques. L'équipe StatInfOmics fait partie des Groupements de Recherche BioInformatique Moléculaire : Modélisation et Méthodologie (BIMMM), Approche Interdisciplinaire de l'Évolution Moléculaire (AIEM) et Génomique Environnementale (GE) et est représentée dans les sociétés savantes SFBI (Société Française de Bioinformatique) et SFdS (Société Française de Statistique).

1.1.2 Méthode de travail et suivi des travaux

Au cours de mon stage, j'ai été initié à l'outil polyvalent Quarto (Allaire et al. (2025)), celui-ci à permis de rendre mes travaux disponibles, sous forme de blog hébergé sur la forge institutionnelle

INRAE, associé à un dépôt Git et en accès restreint aux membres du dépôt, tout en garantissant un système de gestion de version. Pendant toute la durée de mon stage, j'ai été suivie lors de réunions hebdomadaires réunissant mes 3 encadrants. En dehors de ces créneaux, les échanges ont été réalisés via Mattermost, me permettant de demander conseil à tout moment et surmonter d'éventuelles difficultés. L'ensemble des recherches bibliographiques liées à mes travaux ont été regroupées et organisées via l'outil Zotero (Digital Scholarship (2025)). J'ai également pu bénéficier d'un accès aux ressources bibliographiques de l'unité, notamment pour l'ouvrage *R Packages* Wickham et Bryan (2023).

Tout au long du stage, j'ai pu assister aux séminaires d'unité mensuels, aux séminaires d'équipes bimensuels et j'ai pu bénéficier d'un accès à distance à la conférence JOBIM (Journées Ouvertes en Biologie, Informatique et Mathématiques), évènement annuel rassemblant les acteurs de la bioinformatique. J'ai également eu l'occasion de présenter mes travaux à deux occasions : (i) la journée des stagiaires de MaIAGE et (ii) une présentation dans l'équipe GiBBS de GABI.

1.2 Le projet HOLOBIONTS

Le projet emblématique HOLOBIONTS, financé par le PEPR (Programmes et Équipements Prioritaires de Recherche) Agroécologie et Numérique, considère l'holobionte comme une nouvelle échelle biologique pour explorer la diversité génétique et affiner les stratégies de sélection en agroécologie.

Les animaux d'élevage vivent dans des environnements changeants et complexes, et leur microbiote représente un moyen prometteur de moduler les caractéristiques agroécologiques, en tandem avec leur génétique. En effet, les animaux et leur microbiote forment un organisme complexe, appelé holobionte, qui peut être considéré comme l'unité ultime sur laquelle l'évolution et la sélection opèrent (Theis et al. 2016). L'intégration des données du microbiote dans les modèles de prédiction génomique offre la possibilité d'améliorer les prédictions des phénotypes et des valeurs génétiques. Cependant, cela implique de relever le défi d'intégrer des données hétérogènes et parfois incomplètes, tout en tenant compte des interactions complexes entre la génétique de l'hôte et le microbiote. En effet, dans les premiers instants de l'existence d'un mammifère, le contact maternel pendant la mise bas et l'allaitement joue un rôle crucial dans l'établissement de la composition initiale du microbiote. Par la suite, les gènes de l'hôte et les facteurs environnementaux influencent la colonisation, le développement et la fonction du microbiote, ce qui contribue à la formation des phénotypes de l'hôte. En raison de ces interactions complexes entre le génotype, le microbiote et l'environnement, il reste des défis statistiques et informatiques à relever pour intégrer simultanément toutes les informations disponibles au niveau de l'hôte et du microbiote, mais également pour simuler des données réalistes pour l'évaluation comparative des modèles.

La thèse de Solène Pety "Méthodes hologénomiques pour prendre en compte le microbiote de l'hôte dans les évaluations génétiques" s'inscrit dans le workpackage 2 du projet HOLOBIONTS. Ainsi, le cadre de simulation de RITHMS (R Implementation of a Transgenerational Hologenomic Model-based Simulator, Pety et al. (2025)) vise à modéliser des données transgénérationnelles hologénomiques avec pour particularité de prendre en compte la transmission et la modulation du microbiote, de s'appuyer sur des données réelles issues d'une population fondatrice, de définir une architecture génétique et un programme de sélection via le package MoBPS (Pook et al. 2020), et

enfin de permettre la simulation de scenarii variés et biologiquement réalistes.

1.3 Concepts et définitions mobilisés au cours du stage

La méthode implémentée dans RITHMS se réfère à plusieurs concepts biologiques, statistiques ou bioinformatiques importants. Parmi eux, certains ont été mobilisés afin de caractériser les relations entre génétique, microbiote et phénotype. Au niveau des données manipulées, les communautés microbiennes sont décrites à l'aide d'OTUs (Operational Taxonomic Units), ou plus largement par des taxas (espèce, genre, phylum par exemple). En parallèle, la diversité génétique de l'hôte est caractérisée par des SNPs (Single Nucleotid Polymorphism), variation d'une seule paire de bases dans la séquence. A partir de ces SNPs, il est possible d'identifier des régions génomiques associées à certains caractères phénotypiques, appelées QTLs (Quantitative Trait Loci). Deux types de recrutement du microbiote peuvent être définis : (1) la **transmission verticale**, c'est-à-dire de la mère à la descendance (par le biais de la gestation, de la mise bas, ou de l'allaitement) et (2) le **recrutement horizontal**, par le biais de l'environnement ambiant ou par contact avec les congénères.Concernant le phénotype, plusieurs quantités d'intérêts peuvent être ainsi décrites : (1) l'**héritabilité directe** désigne la fraction de la variance phénotypiques expliquée par le génotype (effet génétiques directs sur le phénotype) ; (2) la **microbiabilité** désigne la fraction de la variance phénotype expliquée par le microbiote ; (3) l'**héritabilité totale** qui comprends : (i) l'effet génétique direct et (ii) l'effet génétique indirect expliqué par la fraction du microbiote ayant un effet sur le phénotype et étant sous contrôle génétique.

1.4 Objectifs

Mon stage s'inscrit dans une démarche d'amélioration de RITHMS afin d'étendre son spectre d'application et le rendre plus robuste face aux cas limites. Plusieurs objectifs ont été fixés, me permettant de m'orienter à la fois sur des question scientifiques et sur des aspects techniques propres au développement logiciel.

Le premier objectif est de **vérifier la robustesse de la méthode** via la mise en parallèle des résultats Porcins (jeu de données utilisé dans l'article Pety et al. (2025) présentant la méthode) et des résultats Bovins. Ainsi, nous pouvons évaluer et vérifier la généricité de l'approche sur des données de compositions, structures et dimensions différentes.

Ensuite, le deuxième objectif est de **constituer une version améliorée et plus intuitive du package** dans une partie ingénierie logicielle. Le but est notamment d'implémenter de nouvelles fonctionnalités et des aides à la calibration, d'optimiser le code et également d'enrichir la documentation.

Finalement, une étude de cas sera menée afin d'**illustrer une utilisation concrète de l'outil dans un contexte à la fois écologique et industriel**. Elle permettra d'évaluer *in silico* l'impact d'une augmentation de la température ambiante sur la composition des microbiotes. Cet effet sera exploré via des simulations transgénérationnelles projetées sur une période de 30 ans, permettant d'analyser les dynamiques à long terme.

Matériel & Méthode

2.1 Données utilisées à des fins d'analyses et de développement du package

2.1.1 Données Porcines

Les données Porcines consistent en 1845 taxa et 5000 SNPs caractérisés sur 750 animaux. Elles sont disponibles via l'objet `Deru` du package `RITHMS`. Les animaux ayant fourni les données de génotypes et de microbiotes sont issus du même protocole expérimental. Les porcs de race pure Large White proviennent des entreprises de sélection Nucléus (Le Rheu, France) et Axiom (Azay-sur-Indre, France), tous ont été élevés à la station de phénotypage France Génétique Porc du Rheu (UE3P, INRAE) en 2018.

Microbiote

Les données microbiote du jeu de données de référence pour le développement de RITHMS proviennent du papier (Déru et al. 2022). Les individus suivant le régime conventionnel ont été retenu, ce qui représente un total de 750 porcs. Les échantillons fécaux ont été utilisés pour le séquençage et l'analyse de l'ADN ribosomal 16S. L'ADN microbien a été extrait à l'aide du kit *Quick-DNA Microbe Miniprep* (Zymo Research, Freiburg, Allemagne). Les régions V3-V4 du gène 16S ont été ensuite amplifiées. Le jeu de données final contient 1845 OTUs OTU avec une prévalence supérieure à 5%.

Génotypes

Les génotypes utilisés sont issus de l'étude précédente, (Déru et al. 2020). Les porcs ont été génotypés avec une puce 70 0000 SNPs *GeneSeek GGP Porcine HD* (Neogen) par le laboratoire d'analyse de Gènes Diffusion (Lille, France). Un sous-ensemble des 5000 premiers SNPs a été conservé pour la production des figures de l'article (Pety et al. 2025) et des analyses présentées ici.

2.1.2 Données Bovines

Les données Bovines consistent en 4018 taxa et 32204 SNPs caractérisés sur 750 animaux. Elle sont issues du papier (Pérez-Enciso et al. 2021), et sont décrites dans deux études distinctes.

Microbiote

Les données du microbiote ont été récoltées chez 750 vaches laitières de race Holstein issues de 5 troupeaux commerciaux (Difford et al. 2018), plus précisément au niveau du rumen, le premier estomac. Un fichier pour le microbiote bactérien et un second pour le microbiote archéen sont disponibles, dans notre analyse nous nous concentrerons uniquement sur le microbiote bactérien. L'extraction de l'ADN et l'amplification génique de l'ARNr 16S bactérien, ainsi que le séquençage ont été réalisés par GATC Biotech (Constance, Allemagne). Les amorces génétiques de l'ARNr 16S utilisées concernent les régions variables V1-V3. Environ la moitié des échantillons ont été analysés

avec la plateforme Illumina MiSeq et l'autre moitié avec la plateforme HiSeq par GATC Biotech. Aucune transformation préalable des données n'a été réalisée en amont de l'utilisation par RITHMS, intégrant déjà un processus de formatage des données du microbiote. Le jeu de données final est constitué de 4018 OTUs.

Génotypes

Les génotypes sont tirés du papier (Wallace et al. 2019), ils concernent également des vaches laitières de race Holstein réparties en quatre équipes de recherche (Royaume-Uni, Italie, Suède et Finlande). L'ADN génomique a été extrait et quantifié à partir d'échantillons sanguins pour le génotypage des SNPs. L'ensemble des animaux ont été génotypés sur le Bovine GGP Hd (GeneSeek Genomic Profilers). L'hybridation de l'ADN et l'acquisition des données des puces de génotypage ont été réalisés par la société Neogen selon les protocoles du fabriquant (Illumina). Les données de génotypages sélectionnés sont composées de 32204 SNPs, après filtre sur la base d'une fréquence de l'allèle mineur supérieure à 1% et dont la proportion de génotypes manquant ne dépasse pas 1%. Les données manquantes ont été imputées par la moyenne.

Afin de rendre plus fluide la lecture de se rapport, nous dénommerons les données issues de (Déru et al. 2022) sous le nom de *données Porcines* et celles issues de (Pérez-Enciso et al. 2021) sous le nom de *données Bovines*.

2.2 Processus de simulation hologénomique transgénérationnelle avec RITHMS

Génotypes

Le procédé de simulation des génotypes transgénérationnels de RITHMS s'appuie sur l'implémentation performante du package MoBPS (Pook et al. 2020) qui permet de considérer des schémas d'élevage et de sélection complexes sous différents scenarii d'héritabilité. Ainsi, les générations successives non-chevauchantes et les pedigrees sont générés à partir de données réelles fournies en entrée par l'utilisateur. Par défaut, le nombre de générations simulées est fixé à 5 et le nombre d'individus à 500. Pour chaque génération, la proportion de femelles dans la population (0.5) est appliqué aléatoirement sur les individus et indépendamment de la stratégie de simulation. Le pool de reproducteurs pour la génération suivante est constitué de 30% des femelles et 30% des mâles de la génération courante, choisis aléatoirement ou suivant un critère de sélection. Plusieurs critères de sélection sont disponibles dans RITHMS : (1) pas de sélection (par défaut, équivalent à un tirage aléatoire des individus), (2) valeur génétique directe $BV_d^{(t)}$, (3) valeur génétique médiée par le microbiote $BV_m^{(t)}$, (4) valeur génétique totale $BV_t^{(t)}$, (5) diversité alpha du microbiote basée sur l'indice de Shannon et (6) mélange pondéré entre la diversité du microbiote et la valeur génétique totale.

Microbiotes

Le microbiote d'un individu adulte est partiellement hérité du microbiote maternel par transmission verticale et du microbiote environnemental par héritage horizontal. Ce dernier est ensuite modulé par le génome de l'individu G_i via une matrice d'effet génétique β et de potentiels facteurs environ-

nementaux X_i via une matrice d'effets environnementaux θ. Ainsi, RITHMS propose de modéliser les abondances du microbiote d'un individu i à la génération $t(t = 1, ..., n_{gen})$ selon le modèle suivant :

$$CLR(M_i^{(t)}) = CLR(\lambda M_{d(i)}^{(t-1)} + (1-\lambda)M_{a(i)}^{(t)}) + \theta(X_i^{(t)})^T + \beta G_i^{(t)} + \epsilon_i^{(t)}$$

Avec :

- $M_i^{(t)}$, abondances relatives des taxa chez un individu i ($n_b \times 1$)
- λ, proportion du microbiote hérité par transmission verticale de sa mère avant toute modulation.
- $M_{d(i)}^{(t-1)}$, abondances relatives des taxa de la mère pour un individu i ($n_b \times 1$)
- $M_{a(i)}^{(t)}$, abondances relatives des taxa du microbiote ambiant pour un individu i ($n_b \times 1$)
- θ, effets environnementaux sur les abondances des taxa ($n_b \times k$)
- $X_i^{(t)}$, facteurs environnementaux pour un individu i ($1 \times k$)
- β, effets de la génétique de l'hôte sur les abondances des taxa ($n_b \times n_g$)
- $G_i^{(t)}$, génotype de l'individu i ($n_g \times 1$)
- $\epsilon_i^{(t)} \sim \mathcal{N}(0, \sigma_m^2 I_{n_b})$, bruit blanc gaussien multivarié

La transformation CLR permet de supprimer l'effet de compositionnalité dans les données. En effet, le fait de travailler avec des abondances relatives impose notamment d'avoir toujours une somme constante. La transformation CLR est effectuée avec la fonction `compositions::clr()` et correspond à l'équation suivante : $CLR(x) = log\left(\frac{x}{g(x)}\right)$, avec $g(x)$ qui représente la moyenne géométrique des valeurs de x (Gloor et al. (2017)).

Le microbiote ambiant est modélisé en partant du postula que les individus d'une même génération vivent dans des conditions similaires et sont exposés aux mêmes sources de microorganismes. De plus, on considère également que celui-ci évolue peu au cours des générations et est fortement lié à la composition moyenne du microbiote de la génération précédente notée $\bar{M}^{(t-1)}$. Le microbiote ambiant $M_{a(i)}^{(t)}$ auquel est exposé l'individu i et dans lequel il recrute son microbiote, est issu du mélange pondéré (en proportions $1-\pi$ et π) du microbiote moyen $\bar{M}^{(t-1)}$ et d'un microbiote Dirichlet simulé $M_{r(i)}^{(t)}$ centré autour de ce même microbiote moyen. Le microbiote ambiant modélisé ainsi :

$$M_{a(i)}^{(t)} = \pi \cdot M_{r(i)}^{(t)} + (1-\pi) \cdot \bar{M}^{(t-1)}$$

avec $M_{r(i)}^{(t)}$ le microbiote Dirichlet, $\bar{M}^{(t-1)}$ le microbiote moyen et $\pi \in [0, 1]$ agissant comme facteur de pondération.

Le microbiote Dirichlet permet d'ajouter une variabilité inter-individus lors de la modélisation du microbiote ambiant pour prendre en compte la stochasticité du recrutement horizontal. Formellement, le microbiote Dirichlet est défini par :

$$M_{r(i)}^{(t)} \sim Dir(\alpha 0 \cdot \bar{M}^{(t-1)}).$$

où $\alpha 0$ représente le paramètre de concentration, et est calibré de façon à représenter la dispersion de la population de base. Par défaut ce paramètre est fixé à 25, en adéquation avec le jeu de données `Deru` disponible via lebien package. $\bar{M}^{(t-1)}$ permet de centrer la variabilité apportée par le paramètre de concentration autour du microbiote moyen. La disparité autour de la moyenne varie inversement avec $\alpha 0$.

La modélisation du microbiote ambiant est implémentée dans `RTIHMS::compute_mean_microbiote()` et son détail est disponible en Annexe.

Intégration des données hologénomiques

Le procédé permettant l'intégration des génotypes et du microbiote dans l'obtention des phénotypes repose sur le modèle récursif proposé par Pérez-Enciso et al. (2021) :

$$y^{(t)} = \alpha^T G^{(t)} + \omega^T B^{(t)} + \epsilon_y^{(t)}$$

Avec :

- α le coefficient de régression correspondant aux effets des QTLs sur le phénotype $(1 \times n_g)$
- $G^{(t)}$ les valeurs génotypiques de tous les individus à la génération $t(n_g \times n_{ind})$
- ω le coefficient de regression correspodant aux effets des OTUs sur le phénotype $(1 \times n_b)$
- $B^{(t)} = CLR(M^{(t)})$, les abondances CLR-transformées des OTUs de tous les individus de la génération t $(n_b \times n_{ind})$
- $\epsilon_y^{(t)} \sim \mathcal{N}(0, 1)$ bruit gaussien univarié de variance 1. Par ailleurs, la variance unitaire est utilisée pour calibrer α et ω afin d'obtenir les h_d^2 et b^2 cibles.

Ce modèle prend en charge la possibilité que certains OTUs causaux sur le phénotype, peuvent être sous modulation génétique partielle de l'hôte. Ainsi le génome exerce un effet à la fois direct et indirect (médié par le microbiote) sur le phénotype. Les valeurs initiales $\tilde{\alpha}$ et $\tilde{\omega}$ des coefficients non nuls de α et ω sont échantillonnées comme dans la méthode Simubiome (Pérez-Enciso et al. (2021)) à partir de distributions $\Gamma(0.4, 5)$ et $\Gamma(1.4, 3.8)$ respectivement. Les coefficients de regression α et ω sont ensuite calibrés à partir de $\tilde{\alpha}$ et $\tilde{\omega}$ (par un simple changement d'échelle) sur la population de base afin d'atteindre les valeurs d'héritabilité directe h_d^2 et de microbiabilité b^2 fixées par l'utilisateur. Ils restent ensuite fixes au fil des générations simulées.

La fonction principale du package, `RITHMS::holo_simu()`, réalise l'entiereté du processus automatiquement. Une vue d'ensemble du processus de simulation est disponible en Annexe. L'objet retourné comprends à la fois les données simulées pour chaque génération, les métadonnées associées ainsi que les paramètres de simulation.

2.3 Outils et méthodes d'analyse

En amont et après leur utilisation dans `RITHMS`, les données ont fait l'objet d'une analyse exploratoire afin de mieux comprendre leurs spécificités.

2.3.1 Analyse exploratoire des données génomiques

Une Analyse en Composante Principale (ACP) a été réalisée sur les données génomiques afin d'identifier d'éventuelles structures de la population initiale. Les packages `bigsnpr` (v.1.12.21) et `bigstatsr` (v.1.6.2) (Privé et al. (2018)) ont été utilisés afin de gérer certaines particularités des données génomiques, notamment la présence d'éventuelles régions en déséquilibre de liaison. Les régions en déséquilibre de liaison désignent des régions spécifiques du génome où certains blocs de SNPs sont corrélés entre eux, et ce même à des distances éloignées. Ces régions sont importantes à prendre en compte lors des ACPs, car elles pourraient représenter des composantes principales dominantes. Pour corriger ce biais, j'ai utilisé la fonction `bigsnpr::snp_autoSVD()` qui applique une décomposition en valeurs singulières (SVD) tout en tenant compte des régions LD à longue distance. Pour ce faire, un objet de classe `bigSNP` a été généré à l'aide de la fonction `bigsnpr::snp_fake()`.

Le code détaillé ainsi que la représentation graphique associée sont disponible en Annexe.

2.3.2 Analyse exploratoire des données microbiotes

Structure des données

Une ACP a été réalisée sur les abondances relatives transformées avec la transformation CLR. Celle-ci a nécessité les packages `FactoMineR` (v.2.12) (Lê et al. (2008)) et `factoextra` (v.1.0.7) (Kassambara et Mundt (2016)).

Des analyses MDS (Multidimensional scaling) ont également été réalisées pour comparer les communautés microbiennes observées initialement et celles simulées avec RITHMS. Cette méthode repose sur le calcul d'une matrice de distances entre les échantillons, ce qui permet de mesurer les dissimilarités biologiques. Ici, la distance de Bray-Curtis a été utilisée sur les données d'adondances relatives. Les fonction `vegan::vegdist()` (v.2.7-1) (Oksanen et al. (2025)) et `phyloseq::ordinate()` (v.1.52.0) (McMurdie et Holmes (2013)) ont été utilisées respectivement pour calculer la matrice de distance et pour réaliser l'ordination multidimensionnelle. Le code est disponible en Annexe.

Mesures de diversités α

La diversité α est une mesure de la diversité locale des espèces au sein d'une population ou d'un milieu. Elle permet d'évaluer combiens d'espèces différentes sont présentes, ainsi que la répartition de leurs abondances. Pour l'estimer au sein des microbiotes, plusieurs métriques ont été calculées :

— **L'Indice de Shannon**

L'indice de diversité de Shannon représente la probabilité de rencontrer une espèce au sein d'un échantillon, il s'agit là du calcul d'une entropie. Cette métrique prends en compte l'abondance proportionnelle des espèces et la richesse spécifique (nombre total d'espèces présentes). Le package `phyloseq` utilise le logarithme népérien comme base logarithmique. Soit la formule de l'entropie de Shannon suivante :

$$H' = -\sum_{i=1}^{S} p_i \ln p_i$$

Où S représente le nombre de taxa différents, p_i la proportion relative de l'espèce i et ln le logarithme naturel (base e, `base = exp(1)`). L'entropie de Shannon est interprétée comme telle : si l'échantillon est homogène, c'est-à-dire qu'il ne comporte qu'une seule espèce, alors $H' = 0$, si au contraire les S taxa sont présents en abondances égales alors H' est maximum et vaut $H' = \ln(S)$.

— **L'Indice de Simpson (inverse)**

Cet indice mesure la probabilité que deux individus tirés au hasard dans un échantillon appartiennent à la même espèce. Si l'on considère un tirage avec remise, alors l'indice de Simpson est défini ainsi :

$$D = \sum_{i=1}^{S} p_i^2$$

Où S représente le nombre de taxa différents et p_i la proportion relative de l'espèce i.

Plus D est faible, plus la diversité est grande et moins la communauté est dominée par quelques espèces. Plus l'indice est proche de 1, plus la population est hétérogène. L'indice de Simpson diminue avec la diversité, à l'inverse de l'indice de Shannon ; c'est pourquoi son inverse est souvent privilégié. Afin de faciliter l'interprétation ultérieure des analyses, nous utiliserons D^{-1}.

— **La Richesse Observée**

La richesse observée représente le nombre total de taxa détectés dans un échantillon, et ce indépendamment de leurs abondances. Ainsi son interprétation reste intuitive, plus la richesse est élevée, plus l'échantillon contient des taxa différents.

Ces métriques ont été calculées soit manuellement soit via `phyloseq::estimate_richness()`. `RITHMS` propose également un moyen de les calculer directement en sortie de `RITHMS::holo_simu()` via la fonction `RITHMS::richness_from_abundances_gen()`. Celle-ci réalise les estimations de richesse à partir des données d'abondances. Cela nécessite une transformation des données d'abondances en table de comptage, format utilisé par `phyloseq::phyloseq()`. Un exemple d'utilisation est fourni en Annexe.

2.3.3 Analyses exploratoires pour la calibration de paramètres

Pour permettre la simulation de nouvelles données hologénomiques avec RITHMS, certains paramètres ont dû être calibrés en fonction des données Bovines.

Effets génétiques :

Dans le modèle, on suppose que certains OTUs sont soumis à une modulation génétique de l'hôte (voir section Processus de simulation[…]). $\sigma\beta$ représente le paramètre de variance de ces effets génétiques. La fonction `RITHMS::gen_effect_calibration()` permet d'évaluer l'impact de σ_β sur l'héritabilité des OTUs. Généralement, la majorité des OTUs possèdent une héritabilité de l'ordre de 0.1, avec un maximum de 0.5 (Zang et al. (2022)). Pour les données Bovines, la valeur du paramètre `effect.size` a été ainsi fixé à 0.6. Les résultats de l'analyse sont disponibles en Annexe.

Profondeurs de séquençage :

Afin de pouvoir estimer la diversité du microbiote, les données issues de `RITHMS::holo_simu()` doivent être transformées en comptages bruts via un tirage multinomial, pour mimer le processus d'échantillonnage et éviter de saturer la richesse observée. Pour cela, la profondeur de séquençage des données initiales est calculée et la médiane relevée comme un estimateur robuste du tirage multinomial. Le paramètre `size_rmultinom` a été fixé à 150000 pour les données Bovines. Le code et les métriques statistiques de la distribution des profondeurs de séquençage est disponible en Annexe.

L'ensemble des analyses a été effectué dans RStudio sur une base R version 4.5.1, et ce dans un environnement Linux Ubuntu 24.04.2 LTS. Le détail de l'environnement R utilisé est disponible en Annexe.

2.4 Outils de développement du package R

Le développement du package R a été réalisé grâce à l'outil `fusen` (v.0.7.1) ((Rochette et al. 2025)), complété par `devtools` (v.2.4.5) (Wickham et al. 2022) et `usethis` (v.3.1.0) (Wickham et al. 2025a) pour automatiser certaines tâches. L'enrichissement de la documentation a nécessité l'utilisation de `roxygen2` (v.7.3.2) Wickham et al. (2024) et de `pkgdown` (v.2.1.3) ((Wickham et al. 2025b)) pour le déploiement d'une interface web exclusivement dédiée à la documentation du package. Le package et les fonctionnalités qui l'entourent ont été développés en respectant au mieux les prescriptions officielles du *Comprehensive R Archive Network* (CRAN) (R Core Team 2025) ainsi que les bonnes pratiques énnoncées dans *R Packages* de Hadley Wickham (Wickham et Bryan 2023).

La documentation en ligne est automatiquement déployée avec `pkgdown` par le système d'intégration/déploiement continue de GitHub. Les vignettes disponibles sont construites à partir de fichiers `.Rmd` et entièrement reproductibles.

Le dépôt public du package RITHMS (disponible via ce lien) est en miroir d'un dépôt institutionnel GitLab, la forge INRAE, sur laquelle l'ensemble de la partie développement est effectuée. Mon travail a été réparti sur plusieurs branches afin de faciliter la gestions des modifications : la branche `fix/warnings` (pour résoudre les problèmes rencontrés lors de l'éxécution des workflows d'intégration et déploiements continus), la branche `improve-package-doc` (pour intervenir sur la documentation du package) et la branche `dev` (dédiée exclusivement à l'implémentation de nouvelles fonctions ou à la refactorisation du code).

2.5 Mise en place d'un cas d'étude

Dans un contexte de déréglement climatique, la température ambiante moyenne est une variable d'intérêt à prendre en compte dans les futurs rendements agroéconomiques (Verma et al. (2024)). Son impact direct ou combiné à l'humidité (sous la forme d'un indice température humidité (THI) (Committee on Physiological Effects of Environmental Factors on Animals (1971))) a été mis évidence à plusieurs reprises concernant la qualité du lait (Kim et al. (2022), Ceciliani et al. (2024),

Landi et al. (2024), Tao et al. (2020)), l'immunité (Koch et al. (2023)) et les émissions de méthane (Souza et al. (2023)).

Notre objectif est d'utiliser `RITHMS` et les données Bovines pour étudier l'impact de l'augmentation de la température sur la composition du microbiote. Nous utilisons pour cela les prévisions de température à 30 ans du GIEC (Calvin et al. (2023)).

2.5.1 Données utilisées pour modéliser l'effet de température au cours des 30 prochaines années

Pour calibrer nos effets environnementaux, applicables sur chaque génération simulée par le package RITHMS, nous nous sommes appuyés sur les résultats obtenus par Correia Sales et al. (2021). Les auteurs présentent les mesures d'abondances relatives de 10 phyla pour plusieurs températures : - des conditions thermoneutres : température ambiante de 24°C, un taux d'humidité relative (RH) de 69.2% et THI de 73.4 - des conditions de stress thermique : température ambiante de 34°C, 69.1% de RH et THI de 84.6.

Le THI est calculé tel que décrit dans le papier (Mader et al. 2006) soit : $\text{THI} = [0.8 \times T_{\text{ambiante}}] + [\frac{\text{RH}}{100} \times (T_{\text{ambiante}} - 14.4)] + 46.4$. Pour notre analyse, nous avons utilisé les individus du groupe ayant suivi un régime dit pauvre en énergie (37% carbohydrate (glucides) non-fibreux, 12 Mcal d'énergie métabolisable par kg de matière sèche) et une prise alimentaire *ad libitum*.

Les estimations des températures sur les trentes prochaines années ont été extraites des projections publiées par le GIEC (Calvin et al. (2023)), selon le scénario SSP5-8.5 (le plus pessimiste) au niveau mondial. Celui-ci prévoit une augmentation de +2.574°C en moyenne d'ici 2055 et +3.168°C pour le quantile à 95%, par rapport à la période 1850-1900 .

Afin de mesurer l'écart en termes d'abondances entre les conditions de stress thermique et thermoneutre, nous procédons comme suit, en travaillant en échelle logarithmique.

$$\Delta_{HS-TN} = CLR(HS) - CLR(TN)$$
$$\Leftrightarrow \log(\frac{HS}{\bar{HS}_g}) - \log(\frac{TN}{\bar{TN}_g})$$

Avec $\bar{HS}_g$ et $\bar{TN}_g$, les moyennes géométriques des abondances relatives relevées dans les conditions de stress thermique et thermoneutre respectivement.

Introduction d'une variabilité inter-individus de l'effet température sur les phyla :

Pour introduire une variabilité inter-individus de manière contrôlée, nous avons supposé des variations typiques de $\pm 5\%$ autour de la valeur moyenne $\Delta^{\text{phylum}_i}_{HS-TN}$ décrite dans la littérature. La distribution des $\Delta^{\text{phylum}_i}_{HS-TN}$ a pu être modélisée à partir d'une loi normale telle que :

$$\Delta^{\text{phylum}_i}_{HS-TN} \sim \mathcal{N}\left(\Delta^{\text{phylum}_i}_{HS-TN}, \sigma^{\text{phylum}_i}_{HS-TN}\right)$$

Avec $\sigma^{\text{phylum}_i}_{HS-TN} = \Delta^{\text{phylum}_i}_{HS-TN} \times 0.05$.

Aucune variabilité intra-phylum n'est injectée, tous les OTUs appartenant à un même phylum auront le même Δ chez un individu donné. À ce stade, une matrice de dimensions $taxa \times individus$ est construite. Elle représente une première version de la matrice θX utlisée dans la méthode RITHMS (voir section Processus de simulation avec RITHMS), que nous notons ϕX.

Le calcul des Δ étant basé sur les résultats obtenus dans les conditions expérimentales du papier (Correia Sales et al. 2021), c'est à dire un écart de 10°C, nous avons supposé une linéarité de la réponse à la température et réajusté l'effet afin qu'il soit proportionnel à l'augmentation de température annuelle que l'on souhaite modéliser. La matrice finale θX est donc obtenue par la relation : $\theta X = \kappa \cdot \phi X$, où $\kappa = \Delta T°C/10$ est le coefficient de mise à l'échelle pour une augmentation de température de ΔT°C.

Nous avons mis en place deux stratégies pour calibrer la valeur de κ au cours du temps. Ces deux stratégies répondent à des problématiques biologiques différentes.

2.5.2 Calibration des effets de la température sur le microbiote selon une stratégie discrète, Méthode A

Dans cette modélisation nous supposons que les effets de la températures évoluent de manière discrète au cours de 25 années successives. Nous souhaitons mettre en place des effets relativement forts et spécifiques à chaque phylum tout en prenant en compte des changements de dynamiques au cours du temps. Cette stratégie vient de l'idée que la température va favoriser des phyla différents en fonction des paliers, les microorganismes ayant des conditions de croissance différentes.

On nomme $\theta X_{n \to n+5}$, l'effet appliqué aux abondances transformées CLR de chaque générations entre l'année n et $n + 5$. Un theta est fixé pour une période de 5 ans. Afin d'avoir un effet de la température très marqué, nous choisissons d'utiliser la mesure du quantile à 95% (+3.168°C) relevé pour 2055. Le choix de cette valeur est purement arbitraire étant donné que cette méthode n'a pas pour objectif de simuler une réalité environnementale. Nous retrouvons ci-dessous l'effet cumulatif appliqué au total sur les 5 générations des 5 périodes :

$$5 \cdot \sum_{i=1}^{5} \kappa_i = 5 \cdot \frac{3.168}{10} \cdot \Delta_{+10°C} \approx 5 \cdot 0.3168 \Delta_{+10°C}$$

Afin de répartir cet effet cumulé sur les cinq périodes, une pondération croissante est mise en place pour déduire le coefficient propre à chaque période.

$$\kappa_i = i \times r, \text{ où } r = \frac{2 \times 0.3168}{n(n+1)}$$

Ainsi, les coefficients d'échelle pour chaque période sont les suivants : $[\kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5] \approx [0.0211, 0.0422, 0.0633, 0.0844, 0.1056]$. Et les matrices θX d'effet environnementaux : $[\theta X_1, \theta X_2, \theta X_3, \theta X_4, \theta X_5] = [\kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5] \cdot \phi X$.

Pour simuler notre dynamique et éviter qu'un même groupe soit favorisé pendant toute la simulation, nous faisons varier aléatoirement l'attribution des effets tous les 5 ans. Ceci permet de simuler des plages de températures préférentielles dynamiques pour les phyla (Cristóbal et al. (2013)). Ainsi, tous les 5 ans, la matrice ϕX est construite de façon à réattribuer aléatoirement les Δ_{HS-TN} aux différents phyla.

Cette méthode sera dénommée *méthode A* dans la suite des analyses.

2.5.3 Calibration des effets de la température sur le microbiote selon une stratégie continue linéaire, Méthode B

L'objectif de cette 2ème méthode est de calibrer les effets de la températures sur les 30 prochaines années, de la façon la plus réaliste possible.

Dans cette modélisation nous supposons l'augmentation de la température comme étant linéaire au cours du temps (approximant une tendance exponentielle). L' effet de la température sur le microbiote de chaque génération évoluera de manière continue au cours du temps. Pour une simulation sur 30 années consécutives, soit 30 générations successives non-chevauchantes, nous modélisons l'effet de la température par 30 matrices θX différentes.

Afin de s'appuyer au mieux sur une situation réaliste, nous utilisons cette fois-ci la moyenne du gain de température pour chaque année. L'écart de température ΔT est calculé entre chaque années n et $n+1$, l'effet associé ($\kappa_{n\to n+1}$) est considéré comme une fraction de l'effet $\Delta_{+10°C}$ du papier Correia Sales et al. (2021).

$$\kappa_{n\to n+1} = \frac{\Delta T_{n\to n+1}}{10} \times \Delta_{+10°C}$$

Ici, nous ne simulons pas de plages de températures préférentielles dynamiques selon les phyla. La dynamique de croissance des phyla relevés dans le papier (Correia Sales et al. 2021) entre les conditions thermiques neutres et le stress thermique est conservée pour générer une tendance globale inchangée sur la période de simulation.

Cette méthode sera dénomée *méthode B* dans la suite des analyses.

2.5.4 Implémentation dans RITHMS

La méthode RITHMS intègre déjà un moyen d'introduire un effet environnemental sur les microbiotes générés au fil des générations. Cependant, celle-ci nécessite de fournir l'effet à appliquer sous forme d'une matrice θX unique (de dimensions $taxa \times individus$), et de préciser ensuite les générations sur laquelle l'effet doit être appliqué. Dans notre cas d'étude, les matrices θX sont amenées à évoluer entre chaque génération. Une adaptation de la méthode a donc été mise en place.

Parmis les 10 phyla les plus abondants dans les données Bovines, 8 sont présents dans les analyses réalisées dans (Correia Sales et al. 2021). Leurs abondances moyennes pour les conditions TN et HS sont relevées, et les Δ correspondant sont calculés selon les méthodes citées précédemment. Les 5

(méthode A) ou 30 (méthode B) matrices θX sont ensuites créées, seuls les OTUs ayant été identifiés au niveau phylum du jeu de données Bovin sont conservés, ce qui représente 3880 OTUs. L'ensemble des matrices sont regroupées dans une liste nommée, `theta_list`. Chaque élément de la liste est identifié par `theta[numéro]`, la matrice associée est accessible via `theta_list$theta1$theta` et les différents Δ choisis pour chaque phylum sont disponible via `theta_list$theta1$parameters`. Cette dernière information n'a pas d'intérêt particulier dans le processus de simulation avec RITHMS, mais permet d'avoir accès aux paramètres de construction de chaque matrice θX, notamment avec la méthode A qui inclue une dynamique. Un vecteur nommé est également construit afin d'associer chaque génération à une matrice θX, `theta_gen_asso`.

La fonction principale du package `RITHMS` a été adaptée pour prendre en entrée ces deux nouveaux objets. `RITHMS::holo_simu()` peut donc, en fonction du nouveau microbiote à simuler, appliquer le bon effet environnemental associé à la génération en question. Cette fonctionnalité sera ajoutée à la version publique de `RITHMS` prochainement.

Analyses des compositions microbiennes

Les compositions des microbiotes de chaque génération ont été analysées en utilisant les abondances relatives et les abondances CLR-transformées. Pour les abondances relatives, les OTUs ont été aggrégés au niveau Phylum, puis leurs abondances moyennées sur l'ensemble de la population. De même, la transformation CLR a été réalisée sur les données déjà aggrégées au niveau Phylum.

Résultats

3.1 Développement du package R

Cette partie présente le travail de développement réalisé autour du package `RITHMS`. Excepté le code brut des fonctions fourni au départ et la toute première compilation, j'ai réalisé l'ensemble des améliorations fonctionnelles, des optimisations et de l'enrichissement de la documentation.

3.1.1 Compilation du package

La structure minimale de tout package R comprend : un répertoire `R/`, un fichier `DESCRIPTION` (renseigne les informations concernant les auteurs, les fonctionnalités du package, sa licence d'utilisation, ainsi que les dépendenses requises ou conseillées) et un fichier `NAMESPACE` (permet de différencier les fonctions importées de packages externes ou exportées par celui en développement). D'autres sous dossiers peuvent être ajoutés pour contenir les tests, la documentation et les vignettes.

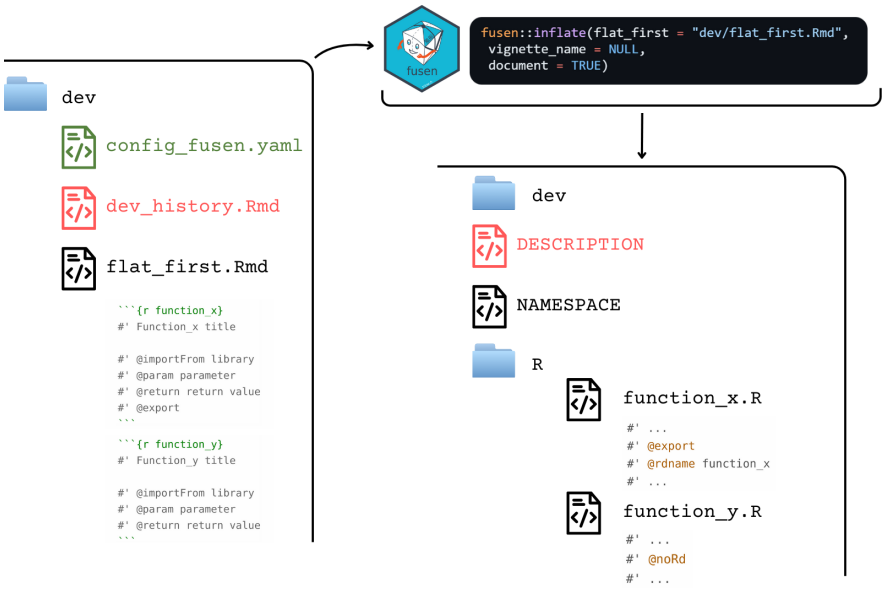


FIGURE 1 : Illustration du processus de compilation du package via `fusen::inflate()`.

L'outil `fusen` permet à partir d'un fichier `.Rmd` de générer l'ensemble de l'arborescence des fichiers et dossiers nécessaires à la création du package. Les fonctions, les tests et les exemples associés sont regroupés dans le même fichier appelé ici `flat_first.Rmd`. Chaque bloc à l'intérieur du `flat_first.Rmd` est identifié distinctement par un en-tête. La compilation via `fusen` permet ensuite de répartir les différentes parties dans les bons fichiers et répertoires du package (Figure 1). Chaque fonction est définie dans son propre bloc. Des balises `roxygen2` sont placées en amont du code. Elles sont essentielles pour décrire le comportement et les arguments de la fonction et générer

la documentation automatiquement. Le premier objectif des balises est de compléter le fichier `NAMESPACE`.

- `@export` indique que la fonction doit être exportée et ajoutée au fichier `NAMESPACE` afin que l'utilisateur final y ait accès.
- `@importFrom` déclare les fonctions issues de package externes nécessaires à la fonction. Elles sont également ajoutées au fichier `NAMESPACE` et permettent de générer les dépendences du package.

Un fichier `dev_history.Rmd` permet de regrouper les métadonnées nécessaires à la compilation du fichier `DESCRIPTION`.

Enfin, lors de la compilation du package, `fusen` exécute également `R CMD CHECK`. Cette fonctionnalité fait partie du processus de validation des packages R. Globalement, la présence des fichiers obligatoires est vérifiée, le code est analysé, et les tests unitaires sont exécutés. Des avertissements peuvent signaler des fonctions non documentées ou des dépendances non déclarées par exemple. A noter que ces vérifications peuvent être effectuées à tout moment avec `devtools::check()`.

3.1.2 Enrichissement de la documentation

3.1.2.1 Documentation du code

Lors de l'écriture des fonctions dans le fichier `flat_first.Rmd`, on retrouve également des balises orientées documentation. Chaque fonction exportée possède un fichier de documentation `.Rd` situé dans le répertoire `man/`. Ces fichiers `.Rd` consituent la documentation officielle du package, accessible via la commande `help([nom_fonction])`. Lors de la compilation, les balises permettent de moduler le contenu des `.Rd` associés. Un exemple complet d'implémentation est présenté en Annexes.

- `@param` identifie chaque argument de la fonction, en précisant son type, sa structure et ses dimensions.
- `@inheritParams` est utilisé pour récupérer les informations de paramètres en commun déjà documentés dans une fonction source. Ceci permet d'éviter la répétition d'informations.
- `@return` et `@value` spécificient ce que retourne la fonction, son type, sa structure et ses dimensions.
- `@seealso` est utilisé pour ajouter un lien vers les fonctions similaires du package.
- `@rdname` permet de préciser le nom du fichier `.Rd` associé à la documentation de la fonction.

Les exemples peuvent êtres implémentés soit dans un bloc à part nommé `examples_[nom_fonction]` ou bien identifié dans le bloc `[nom_fonction]` par la balise `@examples`. Les exemples constituent des cas d'usage minimaux de la fonction et peuvent être executés directement dans l'aide en ligne de R via la section *Run examples*.

`fusen` ne permettant pas pour le moment d'ajouter une documentation pour une fonction non exportée, i.e non visible par l'utilisateur, une adaptation a été nécessaire. J'ai pu implémenter une nouvelle fonction de manière à intervenir directement dans les fichiers `.R` créés par la commande `fusen::inflate()`. Ainsi, les balises `@noRd` sont supprimées, et un appel à `roxygen2::roxygenize()` permet de créer automatiquement les `.Rd` manquants.

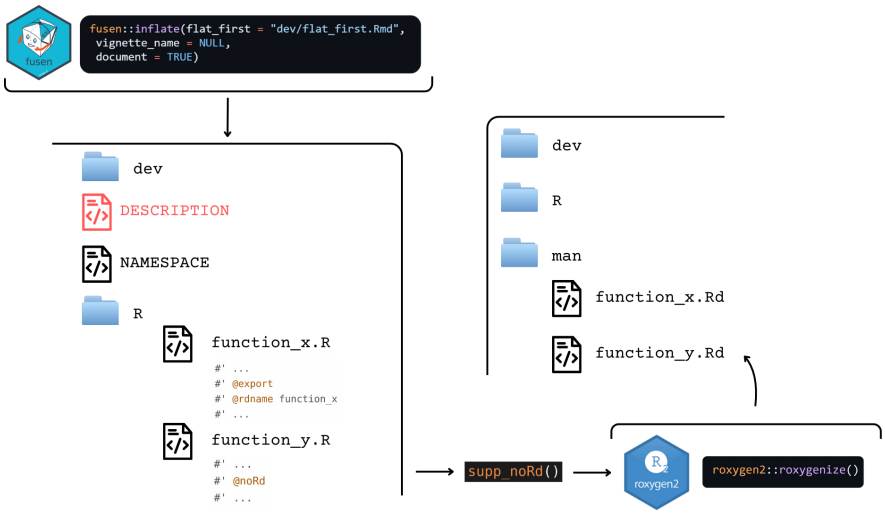


FIGURE 2 : Illustration du processus de compilation du package via `roxygen2::roxygenize()`. Un aperçu de la fonction `supp_noRd()` est disponible en Annexes

3.1.2.2 Documentation pour l'utilisateur des fonctions

La librairie `pkgdown` permet de mettre en place un site web dédié au package (accessible via ce lien). Les instructions d'installation ont été ajoutées en première page, ainsi qu'une prise en main rapide de l'outil. Le contenu de la page d'accueil est défini dans le `README.Rmd`. La commande `devtools::build_readme()` a été utilisée pour compiler le fichier `.Rd` associé, de façon à ne pas recompiler l'entièreté du package si aucune modification n'a été faite par ailleurs.

La structure du site web est décrite dans le fichier `_pkgdown.yml` avec les différentes pages, les liens vers les ressources externes comme le lien vers le dépôt Github du package. C'est dans ce fichier que l'ensemble des fonctions et objets associés au package sont listés, définissant la partie "reference" du site. On retrouve différentes catégories dans cet onglet :

- "Formatting data functions", les fonctions utiles pour préparer et formater les données d'entrée.
- "Taxa assignation function", l'assignation de l'ensemble des taxa à un cluster.
- "Calibration of genetic effect", la fonction diagnostique pour la calibration de l'effet génétique sur le microbiote en fonction des données d'entrées.
- "`holo_simu()`" la fonction principale de l'outil
- "Useful functions relative to `holo_simu()`"
- "Get functions" fonctions utiles pour manipuler les sorties de `holo_simu()`
- "Estimate diversity metrics" pour réaliser les calculs de diversités sur les données microbiote
- "Other little useful functions across the `holo_simu()` process"
- "Dataset" pour introduire le jeu de données Porcines (`Deru`) inclu

Un aperçu du site web peut être généré, grâce à la commande `pkgdown::build_site()`, avant d'envoyer les modifications sur le dépot distant. Ce qui permet de vérifier le bon fonctionnement des onglets et la complétude de la documentation et des exemples.

3.1.2.3 Vignettes d'aide à la prise en main

À partir de la barre de navigation du site, il est également possible d'accéder à plusieurs articles appelés "vignettes". Les vignettes permettent de dédier une page à un sujet ou une fonctionnalité précise. Elles sont compilées en précisant le nom de l'article associé au fichier `.Rmd` dans la commande `fusen::inflate()`. L'ensemble des vignettes peuvent être également compilées en une seule fois via `devtools`.

```
fusen::inflate(flat_file = "dev/calibrate_params.Rmd",
vignette_name = "Calibrate simulation parameters", document = TRUE)

devtools::build_vignettes()
```

Trois vignettes sont actuellement disponibles sur la documentation en ligne :

— "Generate figures" permet à l'utilisateur de générer l'ensemble des figures de l'article (Pety et al. (2025)) à partir du jeu de donnée `Deru`. Celle-ci était déjà existante, mais je suis intervenu au sein des blocs pour harmoniser les thèmes des graphiques.
— "Data importation" a été pensée pour permettre un peu plus de souplesse dans les données d'entrée acceptées par l'outil et permettre à un utilisateur de comprendre le format attendu.
— "Calibrate simulation parameters" a été mise en place pour orienter les utilisateurs dans le choix de certains paramètres, suite aux problèmes rencontrées lors du passage à une autre espèce (voir section Analyses[…]calibration de paramètres).

Contenu de la vignette "Data importation"

Pour permettre aux utilisateurs d'utiliser de nouvelles données dans RITHMS, la vignette "Data importation" détaille les différents formats acceptés. Elle présente les formats standards déjà reconnus par le package (PED/MAP et VCF) et propose une solution pour transformer des jeux de données externes dans un format exploitable. Pour plus de détails se référer à la section (Souplesse pour les données d'entrée).

Contenu de la vignette "Calibrate simulation parameters"

La vignette "Calibrate simulation parameters" présente les étapes diagnostiques à réaliser en amont des simulations avec de nouvelles données. Elle se concentre sur deux paramètres principaux (`effect.size` et `size_rmultinom`), qui peuvent changer beaucoup d'un jeu de données à l'autre.

Le premier paramètre, la taille de l'effet génétique, permet de distribuer l'effet des QTLs à travers le microbiote. La valeur par défaut est optimisée pour être en adéquation avec les données Porcines. La fonction `gen_effect_calibration()` permet aux utilisateurs d'évaluer comment différentes valeurs d'effet génétique peuvent influencer l'héritabilité des taxas (voir section Analyses[…]calibration de paramètres). En plus des graphiques initialement inclus dans cette fonction diagnostique, une version interactive des distributions a été ajoutée. L'utilisation de la librairie `plotly 4.11.0`, permet facilement d'adapter la visualisation aux données et d'explorer les distributions en particulier dans les cas où des densités trop élevées écrasent les autres courbes (se référer à l'Annexe).

Le second paramètre correspond au choix de la taille d'échantillonnage `size` de la fonction `stats::rmultinom()`, utilisé lors des calculs d'indices de diversité. Comme expliqué précédemment, la méthode RITHMS inclue la possibilité de sélectionner selon plusieurs critères dont la diversité du microbiote. Dans la fonction `richness_from_abundances_gen()`, les données de

comptage peuvent être obtenues par un tirage multinomial $\sim \mathcal{M}(N, p)$ où p correspond au vecteur d'abondances relatives et N représente la taille de l'échantillon. Ce dernier nécessite d'être adapté au jeu de donnée utilisé et peut être estimé à partir des profondeurs de séquençage observées. Afin d'obtenir une estimation robuste, la vignette propose le code permettant le calcul des profondeurs de séquençage, le moyen de représenter la distribution de celles-ci et enfin quelques mesures statistiques.

Finalement, la manière dont l'utilisateur doit renseigner les paramètres pour la simulation est rappelée en fin de vignette.

3.1.3 Ajout de fonctionnalités et refactorisation du code

3.1.3.1 Compatibilité avec les formats de données classiques

Ajout de la fonction `transform_geno_into_vcf()`

L'utilisation d'un nouveau jeu de données avec le package `RITHMS` a motivé d'autres interventions sur le code en lui même. En effet, outre l'augmentation des dimensions, il existe une grande diversité de formats pour représenter des informations similaires. Notamment en ce qui concerne les données de génotypes qui peuvent être représentées de multiples façons : codage haplotypique 0/1/2 (nombre de copies de l'allèle de référence), génotypes disjoints par chromosomes ou codage additif par exemple. Dans RITHMS, les données génotypes sont traitées puis modélisées par `MoBPS`. Seul les formats `ped/map` et `vcf` sont pris en charge par la fonction `MoBPS::creating.diploid()`. L'ajout de `RITHMS::transform_geno_into_vcf()` a permis de reconstruire les données sous un format `vcf` en générant des métadonnées factices. Chaque valeur 0, 1 ou 2 est convertie dans le format haplotypique VCF au niveau du champ "GT" : 0/0 homozygote pour l'allèle de référence, 0/1 hétérozygote et 1/1 homozygote pour l'allèle alternatif. Les champs tels que "CHROM", "POS", "ID" sont ajoutés afin de rendre le fichier conforme au standard VCF. L'utilisateur peut spécifier s'il souhaite récupérer le résultat sous forme de `data.frame` ou d'un fichier `.vcf`. Un aperçu de la fonction est disponible en Annexes.

Adaptation des fonctions `read_input_data()` et `generate_founder()`

Suite à ce développement, les fonctions `RITHMS::read_input_data()` et `RITHMS::generate_founder()` ont été modifiées afin de pouvoir prendre en entrée des génotypes au format `ped/map` ou `vcf`. Un argument `file_type` permet de préciser la nature du fichier. L'argument `path_to_genotype` prends en compte le nom du fichier. Dans `RITHMS::generate_founder()`, l'argument `file_type` permet de réaliser des vérifications différentes entre les deux types de format et de faire appel à `MoBPS::creating.diploid()` de la bonne façon.

```
# PED/MAP Genotypes
population <- MoBPS::creating.diploid(dataset = haplo, map = map, verbose=TRUE)

# VCF Genotypes
population <- MoBPS::creating.diploid(vcf = path_to_vcf, verbose=TRUE)
```

3.1.3.2 Nouvelle structure de `founder_object` et mise à jour des données intégrées au package

Nouvelle structure de `founder_object()`

Initialement, `founder_object` était constitué d'une matrice microbiote et d'un attribut "population". L'objet `population`, résultat de `MoBPS::creating.diploid()`, pouvait être accessible via la fonction `base::attr()`. Cette structure étant difficile à maintenir dans l'ensemble du programme, notamment parce que plusieurs opérations supprime les attributs, des modifications ont été apportées. `founder_object` a été redéfini comme une liste nommée, l'accès aux deux objets est direct : `founder_object$microbiote` renvoie la matrice microbiote et `founder_object$population` renvoie l'objet construit par `MoBPS`.

Adaptation du jeu de données intégré et accès simplifié aux données

En parallèle, le jeu de données `Deru` fourni avec le package a été adapté à ce nouveau format. L'objet `founder_object` a été sauvegardé avec la commande `usethis::use_data(founder_object, overwrite = TRUE, compress = "xz")` afin de réduire l'espace mémoire occupé, étant donné la taille importante de l'objet. De même, l'option `LazyData: true` a été ajouté au fichier `DESCRIPTION`, permettant un chargement plus rapide des données. Ainsi, l'objet `founder_object` est directement accessible par l'utilisateur après chargement du package.

3.1.3.3 Nouveau paramètre de simulation pour `holo_simu()`

Le paramètre `size` de la fonction `stats::rmultinom()`, utilisé lors de l'estimation des diversités microbiennes dans `richness_from_abundances_gen()`, doit être calibré selon le jeu de données utilisé. Il a donc été placé en tant que paramètre de simulation d'`holo_simu()`. fixé par défaut à 1000, calibre sur le jeu de données `Deru`. Afin de facilement associer une simulation aux paramètres utilisés lors de l'appel à `holo_simu()`, une liste nommée des paramètres a été ajoutée à l'objet final. `size_rmultinom` en fait partie, ce qui à permis de généraliser le code utilisé dans la vignette "Generate figures" lors des calculs de diversités (*Alpha-diversity remains stable across generations*). Voir en Annexe le code associé au calcul de la diversité.

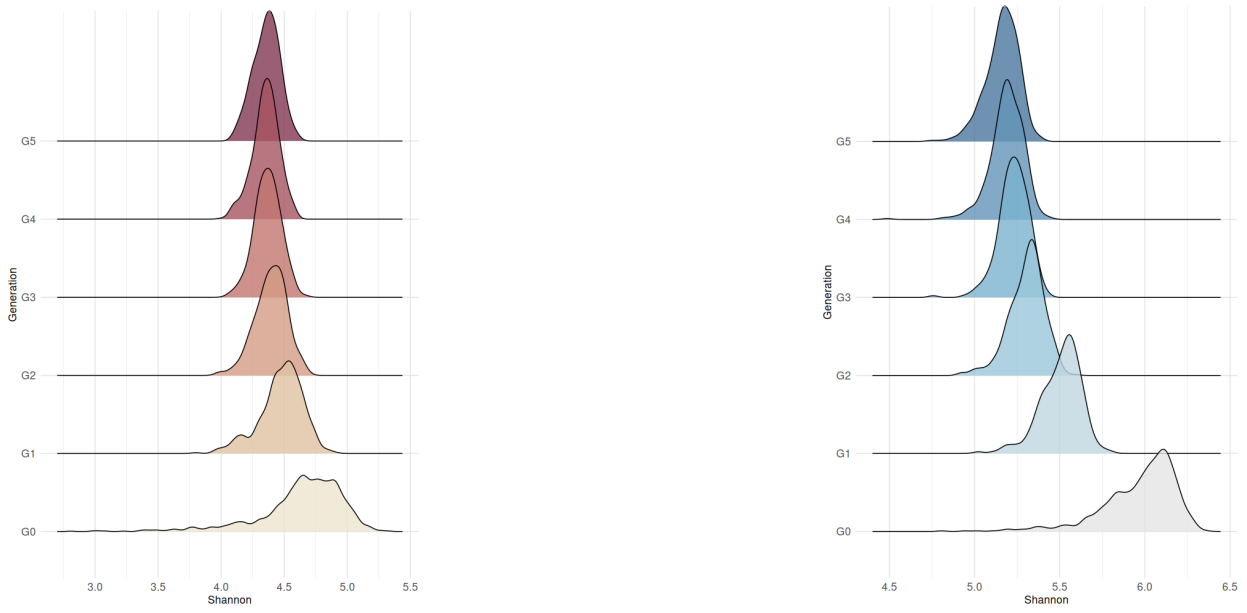
3.2 Mise en situation de RITHMS avec un nouveau jeu de données

Le jeu de données issu de Pérez-Enciso et al. (2021) permet de tester la robustesse du package sur d'autres données que celles utilisées lors du développement. Elles présentent des différences majeures, décrites dans la section Données Bovines, notamment d'un point de vue des dimensions mais aussi des caractéristiques biologiques propres à l'espèce. Une analyse exploratoire préalable aux simulations avec RITHMS a permis de mieux comprendre les spécificités des données.(voir section Analyse exploratoire pour le processus complet). Suite à ces analyses, le paramètre `effect.size` à été fixé à 0.6 (voir en Annexe son impact sur l'héritabilité des taxas)et le paramètre `size.rmultinom` à 150000 (voir en Annexe son impact sur la diversité de Shannon).

Dans en premier temps, l'objectif a été de reproduire les figures présentées dans l'article Pety et al. (2025) afin de comparer la consistance des résultats sur un nouveau jeu de données. L'ensemble des figures produites sont en Annexe. Une première vérification a été de s'assurer que la diversité α se maintienne bien au cours des générations dans un contexte simple sans modulation environnementale et sans force de sélection.

3.2.1 Enjeu : le maintien d'une diversité α

Pour s'intéresser à la diversité, les abondances relatives disponibles à la sortie d'`holo_simu()` sont transformées en comptages à l'aide de la fonction `richness_from_abundances_gen()`. Elle permet ensuite de récupérer les métriques de diversités d'intérêt précédemment décrites.



(a) (b)

FIGURE 3 : Distribution des diversités de Shannon au cours des génération simulées avec RITHMS (G1 à G5). (a)Analyse réalisée sur 1845 OTUs et 50000 SNPs, sur un total de 750 individus, pour les données Porcines. Paramètres de simulation `holo_simu()` : `effect.size=0.3`, `size.rmultinom=1000`. (b)Analyse réalisée sur 4018 OTUs et 32204 SNPs, sur un total de 750 individus, pour les données Bovines. Paramètres de simulation `holo_simu()` : `effect.size=0.6`, `size.rmultinom=150000`.

La représentation des diversités de Shannon (Figure 3) rends compte d'un décalage assez important entre la génération G0 et la génération G1. Ce décalage est davantage marqué avec les données Bovines (Figure 3b) qu'avec les données Porcines (Figure 3a). On note à partir de la génération G1 une stabilisation de la distribution des diversités pour les deux espèces.

Le microbiote d'un individu étant défini comme une contribution pondérée par λ entre le microbiote maternel et le microbiote ambiant(voir section Processus de simulation), une seconde analyse jouant sur ce levier a été réalisée afin de comprendre l'origine cette perte de diversité.(Figure 4).

λ représente la proportion du microbiote maternel transmis à l'individu i et $1 - \lambda$, la proportion du

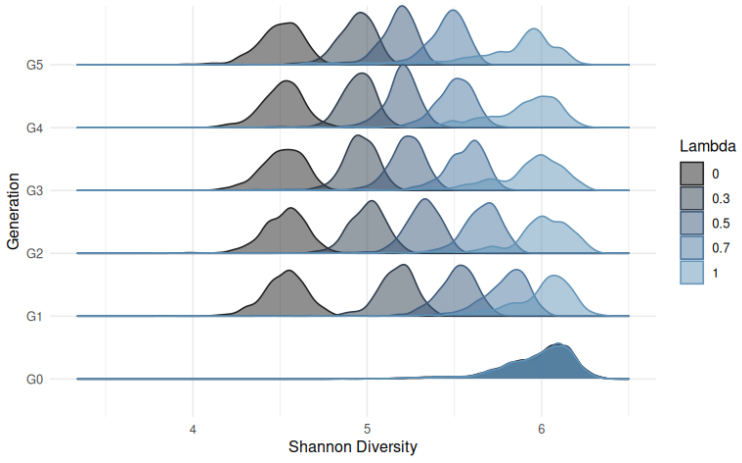


Figure 4 : Densités des diversités de Shannon au cours des générations simulées en fonction du paramètre λ. Analyse réalisée sur 4018 OTUs et 32204 SNPs, sur un total de 750 individus, données Bovines.

microbiote ambiant recruté par l'individu i. Une forte contribution du microbiote maternel ($\lambda \approx 1$) atténue considérablement la perte de diversité observée précédemment entre G0 et G1 (Figure 4), amenant à une diversité de Shannon centrée en 6 sur les 5 générations succédant G0. La diminution de λ, associée à une forte contribution du microbiote ambiant, vient directement agir sur le décalage entre G0 et G1, pouvant diminuer la diversité de Shannon à ≈ 4.5 lorsque $\lambda = 0$. Le paramètre λ constitue en réalité un moyen de modéliser les différentes possibilités de transmission verticales chez les individus. En effet, celui-ci sera d'autant plus important chez les vivipares (Cortes-Macías et al. (2021), Rutayisire et al. (2016)), où les contacts lors de la mise bas sont omniprésents par rapport aux ovipares par exemple (Shterzer et al. (2023)).

Ainsi, la perte de diversité observée semble venir de la construction du microbiote ambiant.

3.2.2 Implication du microbiote ambiant

Rappelons que le microbiote ambiant produit à la génération t, noté $M^{(t)}_{a(i)}$, est issu de la contribution pondérée du microbiote moyen $\bar{M}^{(t-1)}$ et du microbiote Dirichlet simulé $M^{(t)}_{r(i)}$ centré autour de ce même microbiote moyen (voir section Processus de simulation).

Implication du microbiote Dirichlet

La dispersion du microbiote Dirichlet autour du microbiote moyen est contrôlée par le paramètre $\alpha 0$. Dans la figure 6 (Figure 5), nous avons fait varier la valeur de $\alpha 0$ et analysé son impact sur les diversités de Shannon des microbiotes de chaque génération. La diversité de Shannon relevée en génération G1 et G2 semble augmenter avec la valeur de $\alpha 0$. Inversement, l'écart entre la diversité en G0 et celle des générations suivantes diminue avec l'augmentation de $\alpha 0$. Une valeur élevée de $\alpha 0$ permettrait de conserver la diversité de Shannon initiale.

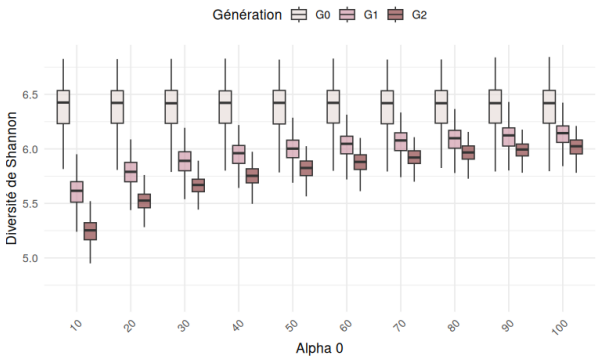


FIGURE 5 : Distribution des diversités de Shannon au cours des générations simulées en fonction du paramètre de concentration de la loi Dirichlet $\alpha 0$. Le paramètre λ utilisé correspond à la valeur par défaut, 0.75 de transmission verticale. Le paramètre π permettant la pondération entre $M^{(t)}_{r(i)}$ (microbiote Dirichlet) et $\bar{M}^{(t-1)}$ (microbiote moyen) correspond également à la valeur par défaut. Analyse réalisée sur 4018 OTUs et 32204 SNPs, sur un total de 750 individus, données Bovines.

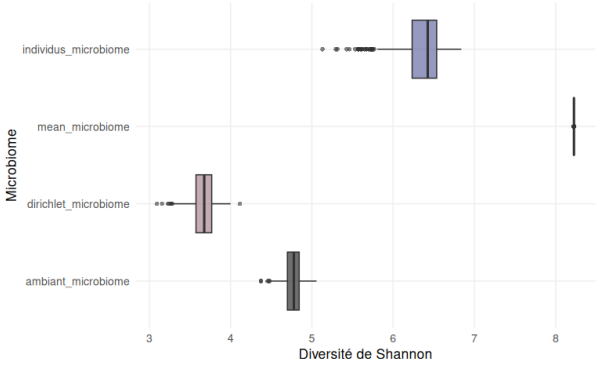


FIGURE 6 : Diversité de Shannon des microbiotes d'intérêt. Le microbiote moyen ("`mean_microbiote`") calculé sur le microbiote global de la population à la génération t ("`individus_microbiote`"). Le microbiote dirichlet ("`dirichlet_microbiote`") centré autour du microbiote moyen avec un paramètre de concentration $\alpha 0$. Enfin, le microbiote ambiant ("`ambiant_microbiote`") construit à partir d'une pondération (π) entre microbiote Dirichlet et microbiote moyen, et constituant (pondéré par λ) du microbiote des individus de la génération $t+1$.

Implication du microbiote moyen

Le second terme qui constitue le microbiote ambiant correspond au microbiote moyen de la population. La figure (Figure 6) regroupe les diversités de tous les microbiotes d'intérêt, permettant de comprendre la contribution de chacun à la diversité : le microbiote des individus de la population t, le microbiote moyen, le microbiote Dirichlet et enfin le microbiote ambiant utilisé dans le calcul du microbiote de la génération $t+1$. Le microbiote moyen est celui qui possède la diversité la plus élevée, > 8. À l'inverse, le microbiote Dirichlet est le plus faiblement diversifié, < 4. Le microbiote ambiant, issu de la pondération entre ces deux derniers possède une diversité d'environ 4.75 qui n'atteint pas celle de la population initiale (≈ 6.5).

Nous avons confirmé que la diversité du microbiote ambiant reste systématiquement inférieure à celle des microbiotes individuels, avec une convergence progressive des microbiotes simulés vers ce profil ambiant au fil des générations (se référer à l'Annexe). Cette vérification permet également de s'assurer de la qualité du microbiote régularisé en G0 et sa cohérence avec le microbiote brut fourni en entrée.

3.3 Etude de cas : contexte de coévolution chez les vaches laitières Holsteins en réponse à une augmentation de la température ambiante

L'objectif de cette étude de cas repose sur l'hypothèse selon laquelle la composition du microbiote d'individus pourraient être amenés à évoluer en parallèle de la température terrestre moyenne.

Les deux expérimentations suivantes ont été réalisées sur les données Bovines, sur un total de 750 individus, 3880 OTUs et 32204 SNPs. Le nombre de taxa sous contrôle génétique a été fixé à 0, ce choix à permis d'évaluer uniquement l'effet environnemental sur la composition du microbiote, celui-ci pouvant entrer en compétition avec la modulation génétique.

3.3.1 Effets d'une augmentation de la température : Méthode A

Afin d'apprécier les effets que peuvent avoir une exposition très marquée nous avons appliqué la méthode A, décrite plus haut (voir section Calibration […] selon une stratégie discrète).

Abondances relatives

Dans un premier temps nous représentons les compositions microbiennes des populations de chaque génération en abondances relatives (Figure 7). En parallèle, la valeur moyenne de l'effet (Δ) appliqué à chaque phylum est représenté. Le barplot est tronqué sur une partie de l'axe y car le reste des barres correspond au phylum majoritaire, les *Bacteroidetes*.

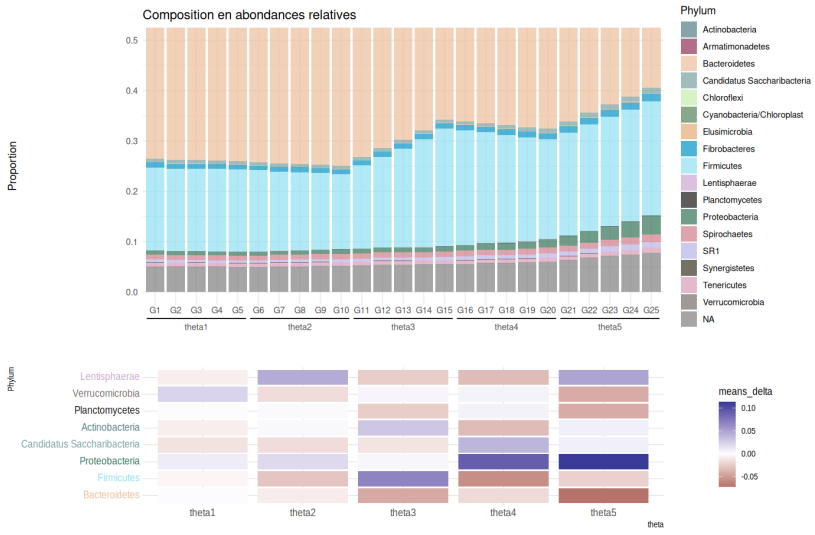


FIGURE 7 : Abondances relatives des microbiotes de 25 générations au niveau phylum, accompagnées des valeurs moyennes de Δ appliqués sur chaque phylum en fonction des θ. Analyse effectuée sur les données Bovines avec 750 individus et 3880 OTUS. Paramètres de simulation : aucun OTUs sous contrôle génétique. Changement des dynamiques de croissance tous les 5 ans.

Une dynamique est observable au cours du temps, notamment au niveau des phyla les plus abondants : *Bacteroidetes* et *Firmicutes*. Même si un peu moins présents, les *Proteobacteria* et les *Candidatus saccharibacteria* semblent eux aussi présenter des changements. Certains autres phyla tels que les *Actinobacteria*, les *Verrumicrobia*, les *Planctomycetes* ainsi que les *Lentisphaerae* sont si peu présents qu'aucun effet sur leurs abondances relatives n'est identifiable. Si l'on compare avec les valeurs moyennes des Δ appliqués sur chaque phylum, on identifie certaines tendances :

— les *Proteobacteria* subissent un effet favorable et augmentent sur l'ensemble des θ.
— les *Bacteroidetes* subissent un effet défavorable pour θ_3 et θ_5, et semblent bien être défavorisés

— les *Firmicutes* subissent un effet défavorable pour θ_1, θ_2 et θ_4, ce qui semble se traduire également sur les abondances relatives. Un effet favorable pour θ_3 semble être lui aussi correspondre à une augmentation des abondances relatives.
— les *Candidatus saccharibacteria* semblent diminuer sur des effet défavorables comme θ_1, θ_2 et θ_3 ; et augmenter sur des effet positifs comme θ_4 et θ_5.

Cependant, certaines tendances contradictoires peuvent êtres observées. Il s'agit par exemple du θ_5 des *Firmicutes*, associé à un Δ négatif, mais avec des abondances relatives qui augmentent sur la période en question. Une observation similaire est observable pour le θ_4 des *Bacteroidetes*.

Abondances CLR-transformées

Pour visualiser plus précisément l'impact des effets environnementaux sur nos compositions microbiennes, les abondances ont été transformées en CLR (Figure 8).

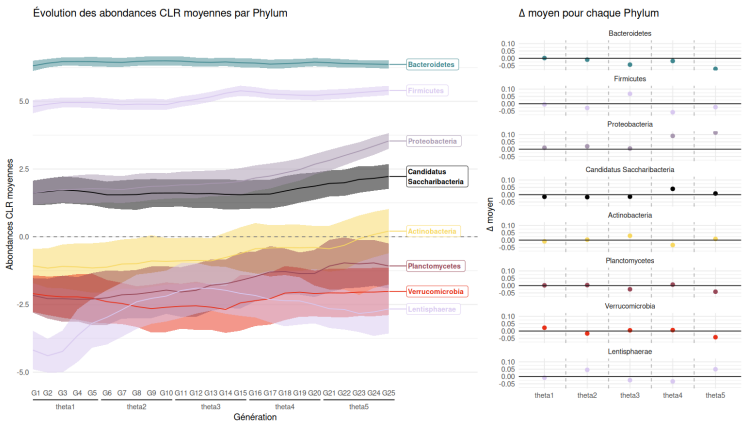


Figure 8 : Abondances CLR-transformées des microbiotes de 25 générations au niveau phylum, accompagnées des valeurs moyennes de Δ appliqués sur chaque phylum en fonction des θ. Analyse effectuée sur les données Bovines avec 750 individus et 3880 OTUS. Paramètre de simulation : aucun OTUs sous contrôle génétique. Changement des dynamiques de croissance tous les 5 ans.

Nous constatons qu'en passant en échelle CLR, des tendances sont identifiables sur les actinobactéries, les *Verrumicrobia* et les *Planctomycetes*, il reste cependant difficile d'établir un lien avec les effets fixés. Aussi, les abondances CLR de deux derniers theta des *Proteobacteria* augmentent (de < 2.5 à > 2.5 en G25), à l'image du Δ correspondant. On peut également identifier un pic au niveau du θ_3 des *Firmicutes*, lui aussi associé à un effet positif. Les abondances des *Bacteroidetes* ne montrent ici pas de tendance particulière contrairement aux observations précédentes (Figure 7). Enfin, les *Lentisphaerae* semblent répondre à l'effet positif en θ_2 et les effets négatifs en θ_3 et θ_4. Cependant, un certain décalage peut être identifié entre la première application de l'effet, et une réponse visuelle au niveau des abondances.

3.3.2 Effets d'une augmentation de la température : Méthode B

La seconde stratégie visait à simuler une augmentation de la température sur les 30 prochaines années de façon la plus réaliste possible. Rappelons qu'aucune inversion de dynamique n'a été introduite dans la croissance des phyla, afin de conserver la tendance observée dans le papier (Correia Sales et al. 2021).

Contrairement à ce qui a été observé dans la méthode A, les abondances relatives des phyla ne montrent pas de réelle tendance identifiable visuellement (Figure 9). En effet, malgré l'ajout d'un

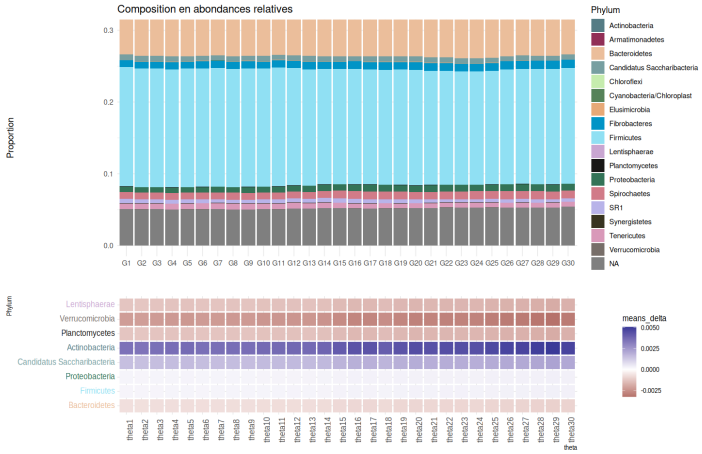


Figure 9 : Abondances relatives des microbiotes de 30 générations au niveau phylum, accompagnées des valeurs moyennes de Δ appliqués sur chaque phylum en fonction des θ. Analyse effectuée sur les données Bovines avec 750 individus et 3880 OTUS. Paramètre de simulation : aucun OTUs sous contrôle génétique, dynamique de croissance inchangée au cours du temps.

effet positifs sur les *Actinobacteria* et les *Candidatus saccharibacteria*, aucune augmentation de leurs abondances relatives n'est relevée. De même, la présence des *Bacteroidetes* semble être stable au cours du temps, en dépit de l'injection d'un Δ négatif.

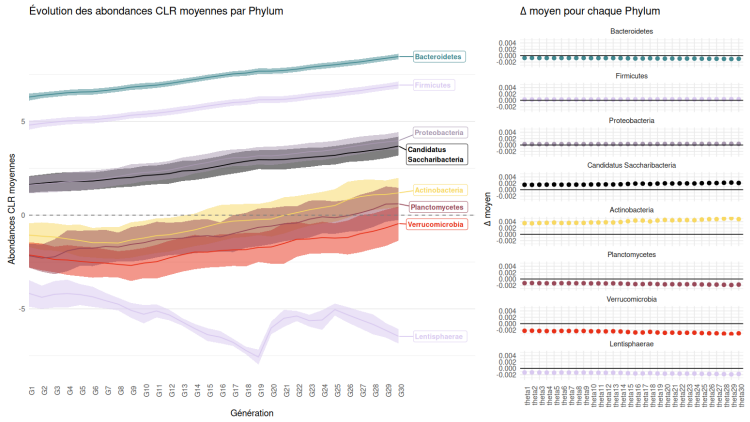


Figure 10 : Abondances CLR-transformées des microbiotes de 30 générations au niveau phylum, accompagnées des valeurs moyennes de Δ appliqués sur chaque phylum en fonction des θ. Analyse effectuée sur les données Bovines avec 750 individus et 3880 OTUS. Paramètres de simulation : aucun OTUs sous contrôle génétique, dynamique de croissance inchangée au cours du temps.

La représentation des résultats en abondances CLR-transformées (Figure 10) fait apparaître des tendances au sein de tous les phyla. Une augmentation des profils est observable chez l'ensemble des phyla excepté les *Lentisphaerae* dont la tendance diminue (de > -5 à < -5 en G30). Les effets négatifs des *Bacteroidetes*, *Planctomycetes* et *Verrumicrobia* ne sont pas repercutés sur les tendances observées. Les *Candidatus saccharibacteria* et les *Actinobacteria* passent de valeurs CLR négatives à des valeurs positives conjointement aux Δ appliqués.

Interprétations & perspectives

4.1 Interprétation des résultats

4.1.1 Le paramètre $\alpha 0$ dans le maintien d'une diversité α

Au cours de l'investigation menée avec les données Bovines, nous avons identifié une perte significative de diversité α entre la population de base, issue des données réelles, et la première population simulée.

Le microbiote des individus à la génération t est construit en 2 étapes, (1) en injectant une variabilité inter-individus via un microbiote Dirichlet $M_{r(i)}^{(t)}$ additionné dans une proportion $1-\pi$ au microbiote moyen $\bar{M}^{(t-1)}$, constituant un recrutement horizontal et (2) en ajoutant dans une proportion λ du microbiote maternel, la partie transmissible du microbiote. Le microbiote ambiant calculé à la première étape joue donc un rôle crucial dans la composition des microbiotes individuels résultants.

En vue d'expliquer cette perte de diversité entre G0 et G1 chez les bovins, nous sommes revenus sur la composition même des jeux de données bruts. Particulièrement en comparant les distributions des prévalences des OTUs avec les données Porcines. La Figure 19 disponible en Annexe, nous montre une répartition bien différente des OTUs sur l'ensemble des échantillons. En effet, alors que la présence des OTUs est très variable d'un échantillon à l'autre dans le jeu de données Porcin, les OTUs dans le jeu de données Bovines sont quasiment omniprésents. A titre comparatif nous relevons près de 599 OTUs dont la prévalence est inférieure à 10% chez les porcs, contre aucun chez les bovins. Cette absence d'OTUs dits "rares" ne correspond pas à la réalité biologique et est révélatrice d'un traitement de données en amont pour filtrer les taxa de faible prévalence. Elle rend également la communauté microbienne de départ assez homogène d'un individu à l'autre.

Le microbiote ambiant est à l'origine d'une perte de diversité (Figure 5), particulièrement lorsque les compositions microbiennes de la population de base sont homogènes comme dans le jeu de données Bovines. Le paramètre de concentration $\alpha 0$ dans le modèle de Dirichlet utilisé lors de la simulation du microbiote ambiant contrôle directement la variabilité inter-individus. Ce paramètre détermine à quel point les profils microbiens seront dispersés autours du microbiote moyen. Ainsi, lorsque α est élevé, les individus tendent à partager une composition similaire, réduisant la variabilité inter-individus et homogénéisant la composition microbienne. A l'inverse, lorsque α est faible, des profils plus hétérogènes sont produits augmentant la variabilité inter-individus. Les MDS illustrant ces observations sont disponibles en Annexes.

L'entropie de Shannon mesure l'incertitude liée au tirage aléatoire d'un OTU. En effet, l'indice de Shannon est calculé à partir des proportions de chaque OTUs dans la population. Ainsi, si les OTUs sont d'abondances équivalentes entre les échantillons, la diversité sera plus élevée, ce qui est le cas dans le jeu de données Bovines à la génération G0. Dans notre simulation du microbiote ambiant, cette situation est possible lorsque le paramètre de concentration $\alpha 0$ est élevé (Figure 5). Les résultats sur les autres indices de diversité α sont présentés en Annexes et confirment les constations faites sur l'indice de Shannon.

Qu'en est-il du rôle du microbiote moyen dans la modélisation du microbiote ambiant ?

Parallèlement, le microbiote moyen jouerait un rôle important dans la conservation de la diversité α au fil des générations. En effet, ce phénomène reposerait sur l'intuition selon laquelle l'entropie du microbiote moyen serait supérieure à la moyenne des entropies des microbiotes individuels. L'inégalité de Jensen illustre cette propriété, le théorème est énnoncé en Annexe. En appliquant cette inégalité à l'entropie de Shannon (concavité) on obtient :

$$H\left(\sum_{i=1}^{n}\lambda_i p_i\right) \geq \sum_{i=1}^{n}\lambda_i H(p_i)$$

Dû au terme logarithmique, l'entropie de Shannon est concave, l'inégalité de Jensen est donc inversée. Ce qui confirme nos observations (Figure 3) et notre intuition selon laquelle la diversité du microbiote moyen est supérieure à la moyenne des diversités individuelles.

4.1.2 L'augmentation de la température, un facteur d'intérêt dans la compositon des microbiotes

Methode A, une dynamique de croissance au sein des phyla

La stratégie d'évolution discrète de la température (Méthode A) met en évidence une réponse assez marquée des phyla les plus abondants (*Bacteroidetes*, *Firmicutes*, *Proteobacteria*). L'application d'effets environnementaux dont la tendance positive ou négative change au cours des générations, permet de visualiser des changements de tendances dans les abondances relatives (Figure 7). Cependant nous avons pu identifier des cas où les abondances relatives ne suivent pas la direction donnée par l'effet appliqué. Par exemple pour le theta 5 associé à un effet négatif chez les *Firmicutes* et les *Bacteroidetes*. Une augmentation des *Firmicutes* est malgré tout observable. L'une des hypothèses pourrait être la présence d'un effet de compensation entre phyla dû à la nature des données de composition. Lorsqu'un phylum subit un effet négatif très marqué (ex. *Bacteroidetes*), cela peut indirectement favoriser la progression d'un autre phylum ; même si ce dernier est lui aussi soumis à effet négatif plus faible. Ce mécanisme pourrait être d'autant plus marqué si l'effet négatif très fort concerne un phylum très abondant. Cette hypothèse illustrerait un mécanisme de compétition et de "redistribution de l'espace microbien", propres aux données d'abondances relatives. Les données transformées CLR ont permis d'appréhender au mieux les tendances réélles (Figure 8). Certains phyla semblent assez réactifs face aux effets positifs ajoutés (*Proteobacteries*, *Firmicutes*). Bien que la tendance chez les *Lentisphaerae* soit visuellement moins corrélée avec l'effet injecté, si l'on compare avec les résultats d'une simulation sans effets environnementaux (Annexes Figure 23), on identifie un changement dans la tendance globale du phylum. Une piste de vérification serait de réaliser un test t de Student pour échantillons appariés sur les données CLR, supposées remplir la condition de normalité. Ce qui permettrait de comparer statistiquement les moyennes des abondances avec et sans les effets température.

Méthode B, tendances visuelles peu marquées

La méthode B, qui simule une augmentation linéaire et réaliste de la température, et ce sans change-

ment de dynamique, montre un profil différent. Aucune tendance ne semble clairement identifiable sur les résultats d'abondances relatives, celles-ci restent comparables visuellement aux résultats obtenus dans une simulation sans effet environnemental (Annexe Figure 22). L'évolution des abondances CLR-transformées est, elle, plus marquée mais n'est pas toujours en adéquation avec les Δ fixés. Le parallèle avec les résultats d'une simulation sans effets environnementaux (Annexes Figure 23) permet cependant de remarquer une atténuation ou une exacerbation des tendances de la condition contrôle. La méthode B permet à l'ensemble des phyla de se retrouver en G30 avec des abondances CLR supérieures à celles relevées en G30 de la condition contrôle. Cette constatation remet en question l'effet des Δ fixés, même si ceux-ci sont négatifs.

4.2 Conclusion

Au cours de mon stage, je me suis appuyée sur la méthode RITHMS pour mener à bien trois objectifs distincts.

Dans un premier temps, j'ai été amenée à évaluer la généricité de la méthode RITHMS sur un jeu de données différent de celui utilisé pour le développement. Les figures de l'article ont été reproduites afin de comparer la consistance des résultats. Cette investigation a permis de mettre en avant les paramètres de la méthode particulièrement sensibles au jeu de données et de proposer une façon de les calibrer au mieux pour refléter les spécificités des données. Dans les analyses menées, la question du maintien de la diversité α au cours des générations et de la construction du microbiote ambiant a été soulevée. Nous avons mis en évidence l'incidence de la structure et de la composition des données sur cette métrique. Cela motive d'autant plus l'intérêt d'utiliser RITHMS sur d'autres jeux de données pour affiner au mieux les recommandations de calibration. Les paramètres par défaut permettent de réaliser des simulations correctes, en considérant toutefois que l'utilisateur puisse les adapter à ses besoins. Selon ses besoins, il lui est possible d'accorder plus d'importance à une variabilité inter-individus ou bien à une homogénéité des microbiotes.

Ensuite, le cas d'étude a permis d'évaluer plusieurs points sur les effets que peut avoir une augmentation de la température ambiante sur la composition du microbiote de vaches laitières. Les simulations réalisées avec une augmentation de la température par palliers et une alternance des effets positifs/négatifs sur les phyla mettent en évidence une forte réactivité du microbiote. Cette réactivité permet notamment de compenser les effets délétères appliqués sur certains phyla pendant une période donnée. Cette dynamique peut être rapprochée de la variabilité saisonnière observée à plusieurs reprises dans les microbiotes de plusieurs espèces (Li et al. (2019), Maurice et al. (2015)), motivée par des changements rapides de conditions environnementales. Un mécanisme de redistribution, propre aux données d'abondances relatives, a été également mis en évidence, suggérant que la diminution d'un phylum dominant peut mécaniquement en favoriser un autre (Gloor et al. (2017)). Dans un scénario plus réaliste d'augmentation continue de la température, les effets appliqués apparaissent trop faibles pour générer des changements nets en abondances relatives, et ce même chez des phyla dominants. Des tendances émergent néanmoins en échelle CLR, des analyses statistiques seraient à prévoir dans l'objectif de caractériser plus précisément ces effets. Enfin, à l'image des conditions expérimentales testées dans le papier Correia Sales et al. (2021), un effet combiné de la température avec d'autres facteurs serait intéressant à mettre en évidence (Williams et al. (2023)). Par exemple, il a été montré plusieurs fois que la prise alimentaire en condition de stress thermique est réduite

(Rhoads et al. (2011), Wheelock et al. (2010)), impactant également la composition du microbiote des animaux (Koch et al. (2024)). Aussi, prendre en compte des changements attendus sur la composition du régime alimentaire en lui même (Baniel et al. (2021)), conséquence d'une altération de la disponibilité des ressources (augmentation des fortes périodes de sécheresse, diminution du taux de précipations par exemple). Ces résultats motivent l'intérêt de considérer à la fois des dynamiques rapides et des tendances progressives sur le long terme, mais également d'étudier les interactions entre différents facteurs : climat, alimentation mais également la génétique de l'hôte implémentée dans l'outil RITHMS. Enfin, l'impact de la tempérture sur le phénotype de l'hôte n'a pas été exploré au cours de nos analyses, ce qui constitue une perspective à prendre en compte dans un exploration future. En effet, des considérations à l'échelle phénotypique sont essentielles à investiguer en raison de l'influence du microbiote sur le phenotype de l'hôte (Zilber-Rosenberg et Rosenberg (2008)).

Dans une partie ingénierie logicielle, j'ai apporté des améliorations fonctionnelles et enrichi la documentation. L'utilisation de RITHMS en conditions réelles a mis en évidence le besoin de construire une documentation riche et accessible à l'utilisateur. Une prise en main de l'outil est disponible en ligne, accompagée d'instructions pour l'utiliser avec de nouvelles données (conversion des données en format compatible, calibration des paramètres spécifiques). Chaque fonction dispose également de sa propre documentation, comprenant des indications sur le format d'entrée attendu et un exemple d'utilisation. Pour rendre l'expérience de l'utilisateur plus intuitive, certaines structures d'objets ont été modifiées et de nouvelles fonctions ajoutées. Sur le plan technique, une attention particulière a été portée à la fiabilité et la reproductibilité du code : (1) grâce à l'utilisation d'un système de gestion de versions, organisé en branches pour assurer la traçabilité des modifications, et (2) grâce aux pipelines d'intégration et de déploiement continu, permettant d'automatiser la vérification du code. Parmi les pistes d'amélioration possibles, l'ajout d'indications dédiées à la calibration des paramètres impliqués dans le maintien de la diversité α serait à prévoir, tout en laissant à l'utilisateur la possibilité d'ajuster au mieux le modèle à ses besoins de simulation. Enfin, un aspect technique important portera sur l'augmentation du taux de couverture des tests unitaires, afin de renforcer la robustesse et la pérennité de l'outil.

Bilan des acquis techniques, méthodologiques et relationnels

Ce stage, effectué sur une période de 6 mois consécutifs, m'a permis d'acquérir de nouveaux savoirs scientifiques, techniques, mais également de façon plus générale au niveau personnel.

Sur le plan technique, j'ai beaucoup progressé dans l'utilisation des systèmes de gestion de versions dans le cadre de projets collaboratifs. Notamment, l'ouverture de nombreuses "merge request" et d' "issue" ont été primordiales pour répondre à des problématiques précises dans le développement du logiciel, en accord avec l'ensemble des collaborateurs. J'ai pu découvrir également les outils d'intégration et de déploiement continus dans le but d'automatiser et fiabiliser le processus de développement du logiciel. Ces deux points constituent des atouts essentiels pour me permettre d'appréhender au mieux mes projets futurs en ingénierie logicielle et déploiement d'outils.

Aussi, mes compétences en bioinformatique ont été développées, en particulier sur les aspects liés aux formats de données souvent complexes dont il est important de maîtriser les secrets pour mener à bien des analyses de qualités. J'ai pu cerner d'autant plus toute l'importance de connaître les données avec lesquelles on travaille, d'explorer leurs spécificités et leur structure avant d'entamer quelconque analyse ou transformation. Par ailleurs, les manipulations des données, les analyses et les représentations graphiques réalisées avec R, m'ont permis d'adapter mon raisonnement algorithmique et de programmation aux spécificités de ce langage. Initialement avec un raisonnement orienté Python, j'ai mis en place des pratiques plus efficaces en R, en limitant l'usage de certains procédés en faveur d'une manipulation vectorielle des éléments. J'ai ainsi gagné en adaptabilité d'un point de vue logique de programmation.

Cette expérience m'a donné l'opportunité de plonger dans un environnement humainement et scientifiquement riche où coexiste des profils allant de l'informatique aux statistiques, en passant par la biologie et parfois même la chimie. J'ai ainsi découvert une grande communauté de bioinformaticiens, et constaté son spectre à l'échelle nationale lors de la conférence JOBIM. Enfin, la participation aux séminaires internes et externes, aux MaIAGE Scientific Days, et à la célébration des 10 ans de MaIAGE, m'a donné l'occasion de m'intégrer à la grande famille que représente une unité de recherche.

Enfin, la diversité des missions que j'ai réalisées autour du package RITHMS, m'a permis d'affiner mes envies professionnelles et mes préférences entre les aspects logiciels, statistiques ou biologiques.

Annexes

A Processus de simulation avec RITHMS

A.1 Modélisation du microbiote ambiant

```
compute_mean_microbiote <- function(microbiote, dir = F,
                                    n_ind = NULL, ao, mix.params){
  mean_microbiote <- rowMeans(microbiote)
  if(dir){
    stopif(is.null(n_ind))
    dir_mean <- rdirichlet(n_ind,as.numeric(mean_microbiote)*ao) %>% t()
    mix_mean <- (mix.params[1] * dir_mean + mix.params[2] *
                 matrix(mean_microbiote,nrow=nrow(dir_mean),
                        ncol=ncol(dir_mean),byrow=F))
    attr(mix_mean, "dirichlet") <- dir_mean
    return(mix_mean)
  }else{
    mean_microbiote
  }
}
```

A.2 Algorithme complet de la méthode

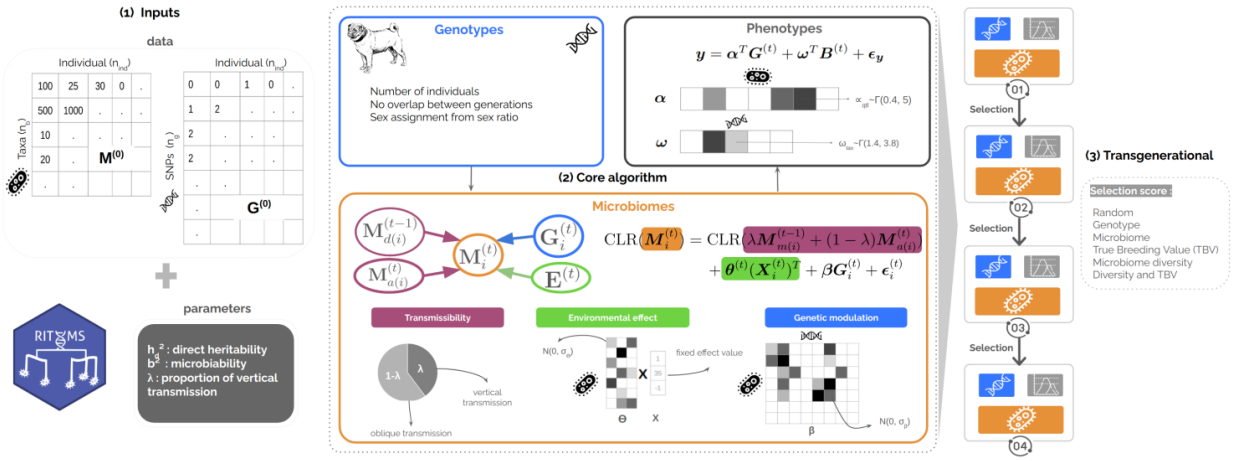


FIGURE 1 : Illustration de la méthode tirée du papier Pety et al. (2025). (1) Les données fournies par l'utilisateur comprennent les abondances relatives du microbiote et génotypes (codés 0/1/2) ainsi que les paramètres suivants : héritabilité directe h^2_d, microbiabilité b^2 et λ, qui module le rapport de transmission verticale/horizontale. (2) Pour chaque génération simulée t, les génotypes et le pedigree sont générés à l'aide du package `MoBPS` (Pook et al. (2020)). Le microbiote est ensuite construit en combinant en premier le microbiote maternel et le microbiote ambiant dans les proportions λ et $1-\lambda$ respectivement, puis en appliquant une modulation génétique et éventuellement environnementale. Les génotypes et le microbiote sont ensuite intégrés pour simules les phénotypes de la génération à l'aide du modèle récursif de Pérez-Enciso et al. (2021). (3) Pour passer à la génération suivante, 30% des mâles et 30% des femelles sont sélectionnés, soit de manière aléatoire, soit selon un indice de sélection choisi par l'utilisateur.

B Analyses exploratoires

B.1 ACP Génotypes

```
library(bigsnpr)

gen <- t(gen)
nb_snp <- ncol(gen)
nb_ind <- nrow(gen)

snp_info <- data.frame(
  chromosome = rep(1, nb_snp),
  physical.pos = seq(1e6, by = 1000, length.out = nb_snp),
  alleles = rep("A/G", nb_snp),
  rsid = paste0("rs", 1:nb_snp)
)

snp <- snp_fake(n = nb_ind, m = nb_snp)

snp$genotypes[] <- gen

snp$map$chromosome <- snp_info$chromosome
snp$map$physical.pos <- snp_info$physical.pos
snp$map$allele1 <- "A"
snp$map$allele2 <- "G"
snp$map$rsid <- snp_info$rsid

acp <- snp_autoSVD(snp$genotypes, snp$map$chromosome, snp$map$physical.pos)

plot(acp$u[,1], acp$u[,2], xlab = "PC1", ylab = "PC2", main = "ACP bigsnpr")
```

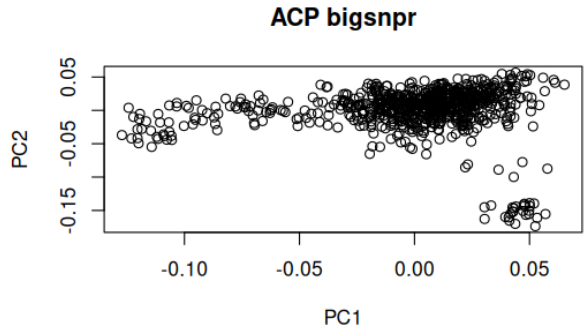


FIGURE 2 : ACP des génotypes sur données Bovines. Analyse réalisée sur 32204 SNPs et 750 individus.

B.2 MDS Bray-Curtis

```
microbiote <- fread("./data/journal.pgen.1007580.s005.txt")
M0 <- microbiote |> column_to_rownames("sample_id")
M0_abund <- apply(t(M0), 2, \(x) x/sum(x))

dist = "bray"

dist_mat <- vegdist(x = M0_abund,
                    method = dist)
```

```
physeq <- phyloseq(
  otu_table(MO_abund, taxa_are_rows = TRUE))

ord <- ordinate(physeq, method = "MDS", distance = dist_mat)

p <- plot_ordination(physeq, ord) +
    geom_point(size = 1, alpha = 0.8) +
    theme_bw()

p
```

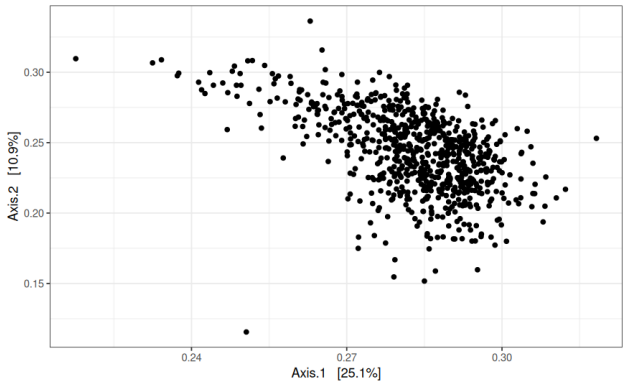


FIGURE 3 : MDS Bray-Curtis du mirobiote des données Bovines. Analyse réalisée sur les 4018 OTUs, utilisation de la distance de Bray-curtis.

B.3 Calculs de diversités α

```
diversities <- generations_simu[-c(1,2)] %>%
  map(get_microbiotes) %>%
  map(richness_from_abundances_gen, size_rmultinom = generations_simu$parameters$size_rmultinom) %>%
  bind_rows(.id = "Generation")

diversity_brut <- phyloseq(otu_table(founder_object$microbiote, taxa_are_rows = FALSE)) %>%
  estimate_richness(measures = measures)
diversity_brut$Generation <- "G0 brut"
```

B.4 Diagnostic de l'effet génétique sur l'héritabilité des taxa

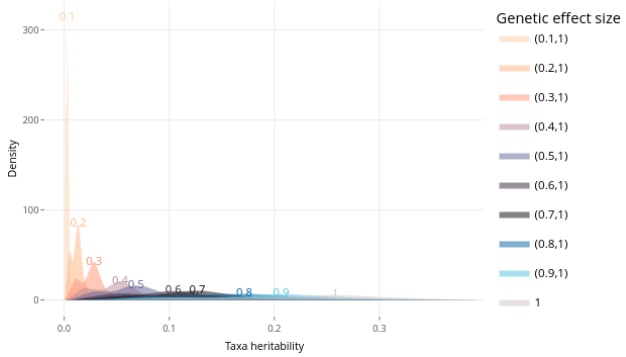


FIGURE 4 : Distribution de l'héritabilité des taxa en fonction de la taille de l'effet génétique choisit.

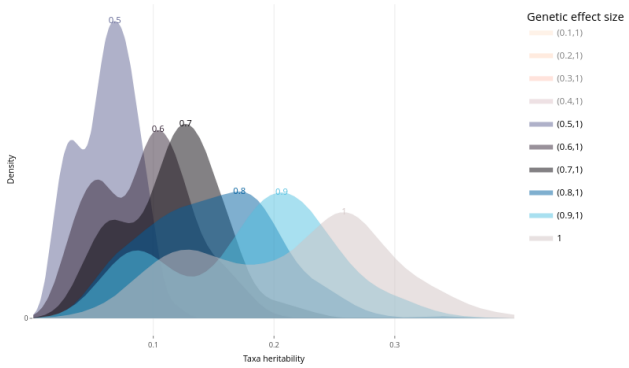


FIGURE 5 : Distribution de l'héritabilité des taxa en fonction de la taille de l'effet génétique choisit. Limité au effets génétiques 0.5, 0.6, 0.7, 0.8, 0.9 et 1. Le package `plotly` permet directement d'interagir avec la visualisation et de sélectionner les distributions d'intérêt.

B.5 Estimation du paramètre `size` de `rmultinom()`

```
microbiote <- fread("./data/journal.pgen.1007580.s005.txt")
microbiote <- microbiote |> column_to_rownames("sample_id")

sample_depth <- rowSums(microbiote)

summary(sample_depth)
    # Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
    #  398    82899   153100   201974   290260 11161859
```

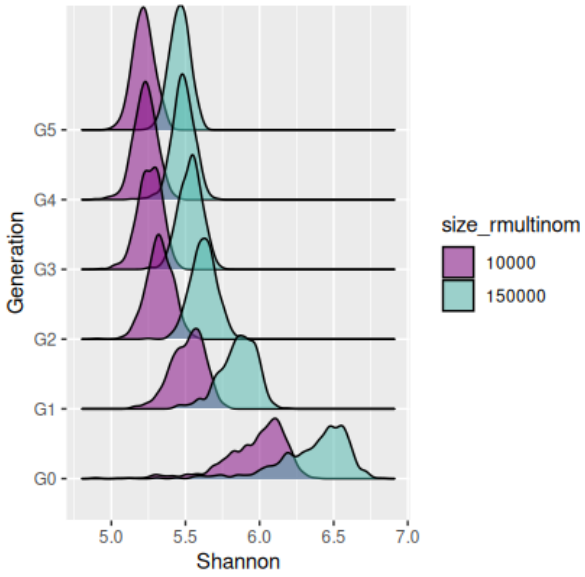


FIGURE 6 : Diversité de Shannon en fonction du paramètre `size_rmultinom`. Analyse réalisée sur les données Bovines, 4018 OTUs et 750 individus. Comparaison entre la valeur par défaut `1000` et `150000` déduite à partir des profondeurs de séquençage.

C Compilation du package R

C.1 Implémentation des éléments relatifs à la documentation d'une fonction

```
#{r function-replace_zero}

#' Useful function to extract microbiotes from RITHMS output
#'
#' The gets functions use the output of [holo_simu()] to extract the information
#' of interest from a given generation.
#' `get_microbiotes` extract the microbiote abundance matrix from a
#' generation object, with or without CLR transformation or transposition.
#'
#' @importFrom compositions clr
#' @importFrom purrr pluck
#' @importFrom tibble as_tibble
#' @importFrom magrittr %>%
#'
#' @param data List corresponding to one generation, as returned by
#' `holo_simu()`. Containing simulation output.
#' @param transpose Logical; if `TRUE`, transpose the microbiote matrix
#' (OTUs in rows, individuals in columns).
#' @param CLR Logical; if `TRUE`, applies a CLR transformation to the abundance
#' data. This transformation requires `transpose = TRUE`.
#'
#' @return A `data.frame`containing the microbiote abundances of individuals.
#' Default, individuals are in rows and OTUs in columns. Change `transpose`
#' parameter if needed.
#'
#' @seealso [get_mean_phenotypes()], [get_phenotypes_value()],
#' [get_om_beta_g()], [get_selected_ind()], [get_phenotypes()]
#'
#' @rdname get_microbiotes
#' @export
get_microbiotes <- function(data, transpose = F, CLR = F) {
  if(transpose){
    if(CLR){
      return(data |> pluck("microbiote") |> t() |> replace_zero() |>
      clr() |> as.data.frame())
    }else{
      return(data |> pluck("microbiote") |> t() |> as.data.frame())
    }
  }
  data |> pluck("microbiote") |> as.data.frame()
}
```

C.2 Fonction `supp_noRd()`

```
supp_noRd <- function(directory = "R") {
  if (!dir.exists(directory)) {
    stop("Le répertoire spécifié n'existe pas.")
  }

  r_files <- list.files(directory, pattern = "\\.R$", full.names = TRUE)

  if (length(r_files) == 0) {
    message("Aucun fichier .R trouvé dans le répertoire.")
  }

  for (r_file in r_files) {
    lines <- readLines(r_file)
```

```
    lines <- lines[!grepl("@noRd", lines)]

    writeLines(lines, r_file)
    message(paste("Le fichier", r_file, "a été mis à jour sans @noRd."))
  }
}

#supp_noRd("path/to/R")
```

D Ajout de fonctionnalités

D.1 Fonction `transform_geno_into_vcf()`

Eventuellement penser à extraire la mini fonction `convert_to_haplo()`

```
#' Convert a 0/1/2 genotype matrix into a VCF-like format
#'
#' This function converts a genotype matrix encoded as 0,1,2 into a VCF-like format.
#' It can either return the VCF content as a `data.frame`, write it to a .vcf file or do
#' both.
#'
#' @param geno_matrix A matrix of genotypes with values 0, 1, 2. SNPs are in rows and
#' individuals in columns.
#' @param output_type A character, that specifies the output type. Choose between `"file"`
#' (write `.vcf`), `"dataframe"` (return `data.frame`), or `"both"` (do both).
#' @param output_path A character string that specifies the output path. Requiered if
#' `output_type` is "file" or "both".
#'
#' @return A `data.frame` if `output_type = "dataframe" or "both"`, or just write the
#' `.vcf` file if `output_type = "file" or "both"`.
#'
#' @rdname transform_geno_into_vcf
#' @export

transform_geno_into_vcf <- function(geno_matrix,
                                    output_type = c("file", "dataframe", "both"),
                                    output_path = NULL){

    n_ind = ncol(geno_matrix)
    n_snp = nrow(geno_matrix)

    convert_to_haplo <- function(x) {
      if(is.na(x)) return(".")
      if (x == 0) return("0/0")
      if (x == 1) return("0/1")
      if (x == 2) return("1/1")
    }

    geno_matrix_haplo <- apply(geno_matrix, c(1,2), convert_to_haplo)

    vcf_snps <- data.frame(
      CHROM = "1",
      POS = 1:n_snp,
      ID = paste0("rs", 1:n_snp),
      REF = "A", ALT = "T",
      QUAL = ".", FILTER = ".", INFO = ".", FORMAT = "GT",
      geno_matrix_haplo
    )

    if(output_type %in% c("file", "both")){
      write.table(vcf_snps, file = output_path, sep = "\t", quote = FALSE,
                  row.names = TRUE, col.names = TRUE)
```

```
    } else if(output_type %in% c("dataframe", "both")){
      return(vcf_snps)
    } else{
      print("The output type must be one of 'file' or 'dataframe'.")
    }
}
```

E Utilisation d'un nouveau jeu de données

E.1 Figures de l'articles reproduites avec les données Bovines

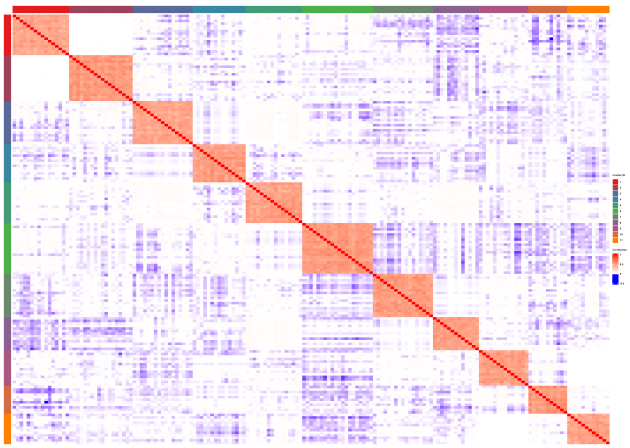


FIGURE 7 : Matrice de corrélation par paires des abondances des OTUS. Les abondances ont été simulées en supposant que tous les taxa sont sous contrôle génétique et répartis en cinq groupes. Analyse réalisée sur le jeu de données Bovines, 750 individus - 4018 OTUS - 32204 SNPs.

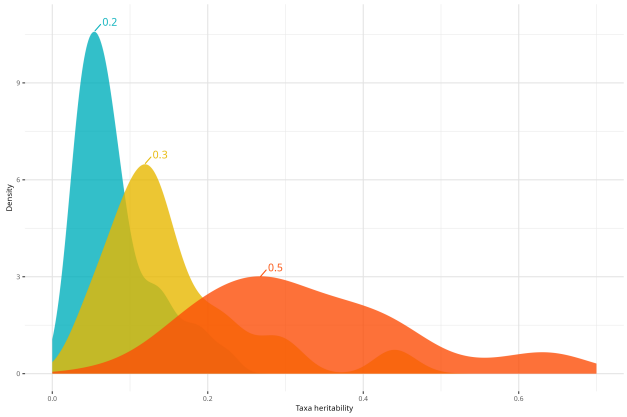


FIGURE 8 : Densité de la distribution de l'héritabilité des taxa en fonction de la taille d'effet génétique. Analyse réalisée sur le jeu de données Bovines, 750 individus - 4018 OTUS - 32204 SNPs.

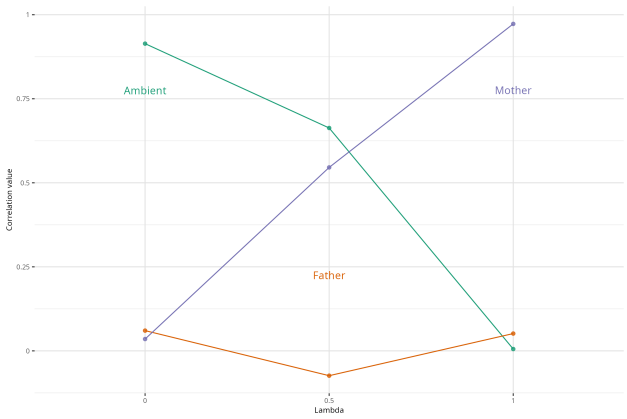


FIGURE 9 : Corrélation entre la diversité α de la progéniture (G2) avec celle de sa mère, de son père ou du microbiote ambiant pour des valeurs croissante de λ. Les corrélation sont calculées à partir d'une population de 500 progénitures. Analyse réalisée sur le jeu de données Bovines, 750 individus - 4018 OTUS - 32204 SNPs.

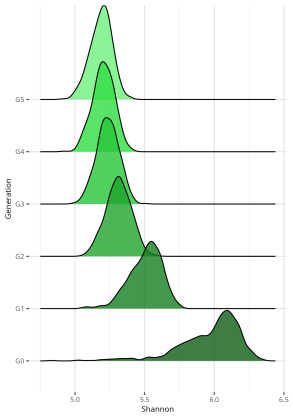


FIGURE 10 : Distribution des valeurs de diversité α sur 5 générations simulées en l'absence de sélection et de filtres environnementaux. Analyse réalisée sur le jeu de données Bovines, 750 individus - 4018 OTUS - 32204 SNPs.

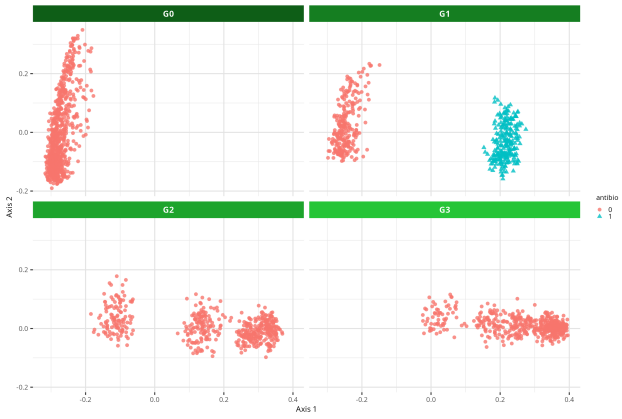


FIGURE 11 : MDS Bray-Curtis des abondances microbiennes. La moitié des individus en G1 sont soumis à un traitement antibiotique sporadique. Analyse réalisée sur le jeu de données Bovines, 750 individus - 4018 OTUS - 32204 SNPs.

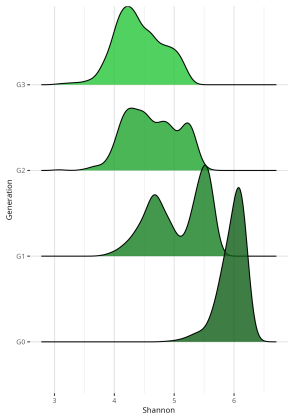


FIGURE 12 : Distribution des valeurs de diversité α avant (G0), pendant (G1) et après (G2 à G3) un traitement antibiotique sporadique.

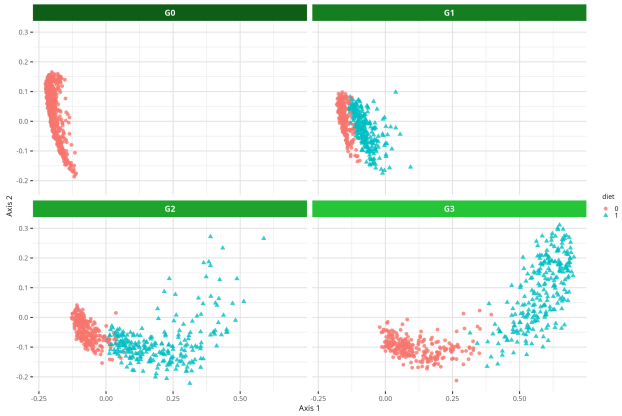


FIGURE 13 : MDS Bray-Curtis des abondances microbiennes. La moitié des individus à partir de G1 sont soumis à un régime alimentaire favorisant deux groupes de taxa. Analyse réalisée sur le jeu de données Bovines, 750 individus - 4018 OTUS - 32204 SNPs.

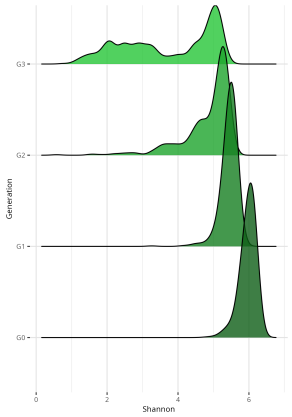


FIGURE 14 : Distribution des valeurs de diversité α avant (G0) et pendant (G1 à G3) une intervention alimentaire soutenue. Analyse réalisée sur le jeu de données Bovines, 750 individus - 4018 OTUS - 32204 SNPs.

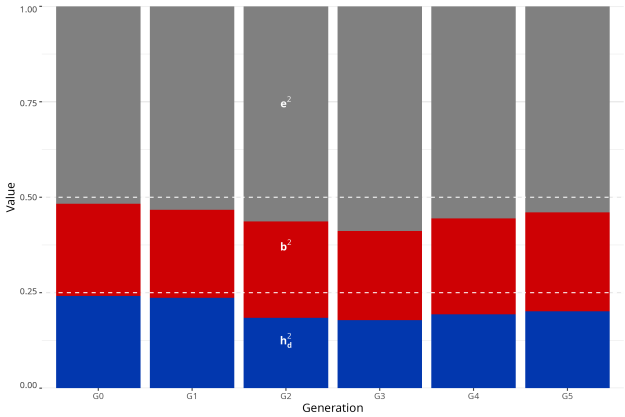


FIGURE 15 : Héritabilité directe et microbiabilité observées dans un scénario sous sélection aléatoire et $h_d^2 = b^2 = 0.25$. Analyse réalisée sur le jeu de données Bovines, 750 individus - 4018 OTUS - 32204 SNPs.

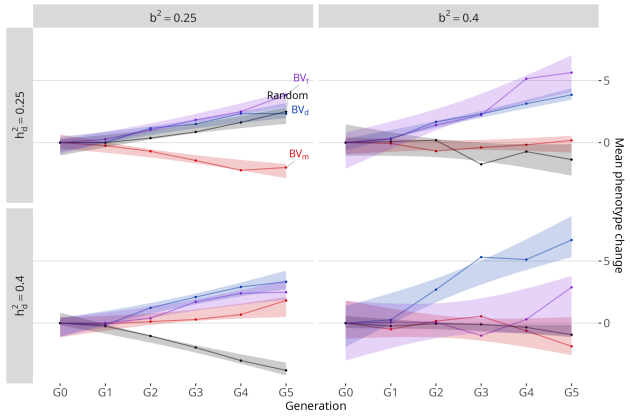


FIGURE 16 : Changement phénotypiques moyens selon diverses valeurs d'héritabilité directe et de microbiabilité, et de stratégies de sélection. BV_d valeurs de sélection directe, BV_m valeurs de sélection du microbiote et BV_t valeurs de sélection totales. Analyse réalisée sur le jeu de données Bovines, 750 individus - 4018 OTUS - 32204 SNPs, avec $\lambda = 0.1$

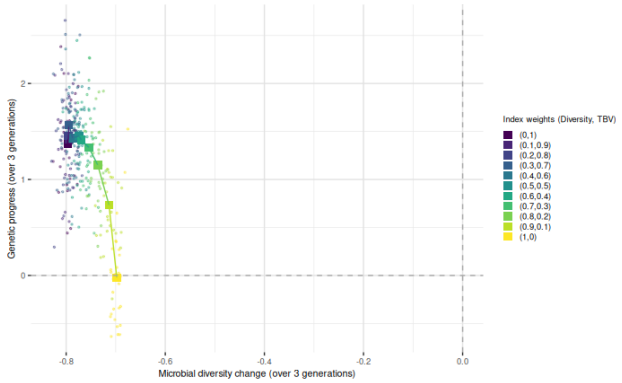


FIGURE 17 : Exploration guidée par simulation de l'indice de sélection mixte. Le phénotype moyen et la diversité microbienne évoluent de la population de base (G0) à G5 en fonction w_{div}. Les moyennes calculées sont représentées par les carrés. Analyse réalisée sur le jeu de données Bovines, 750 individus - 4018 OTUS - 32204 SNPs.

E.2 Microbiotes ambiants et individuels sur 5 générations, porcins et bovins

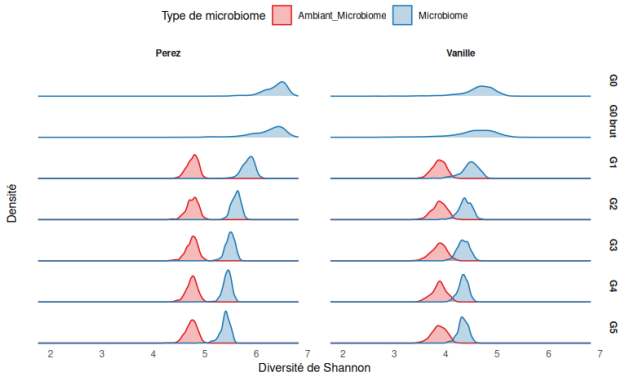


FIGURE 18 : Diversité de Shannon des microbiotes individuels ("`microbiote`") et du microbiote ambiant ("`Ambiant_microbiote`") associé au cours des générations. Le microbiote dit "brut" à également été représenté, il est calculé sur le jeu de données initial avant toute "régularisation" par le package RITHMS. Exceptés les paramètres `size_effect` et `size_rmultinom` spécifiques à chaque jeu de données, l'ensemble des valeurs ont été laissées par défaut. Analyse réalisée sur 4018 OTUs et 32204 SNPs, sur un total de 750 individus, données Bovines ; sur 1845 OTUs et 5000 SNPs, sur un total de 780 individus, données Porcines.

E.3 Prévalence des OTUs

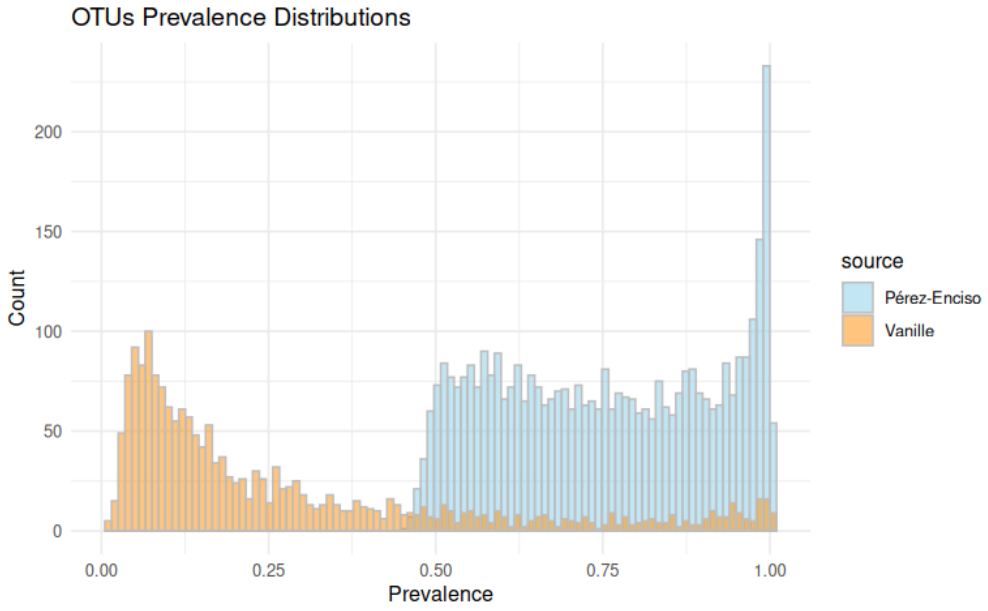


FIGURE 19 : Distribution de la prévalence des OTUs sur les jeux de données Porcins et Bovins. Analyse réalisée chez les porcins ("Vanille") sur 750 individus et 1845 OTUs. Analyse réalisée chez les bovins ("Pérez-Enciso") sur 750 individus et 4018 OTUs.

E.4 Impact du paramètre $\alpha 0$ sur la variabilité inter-individus

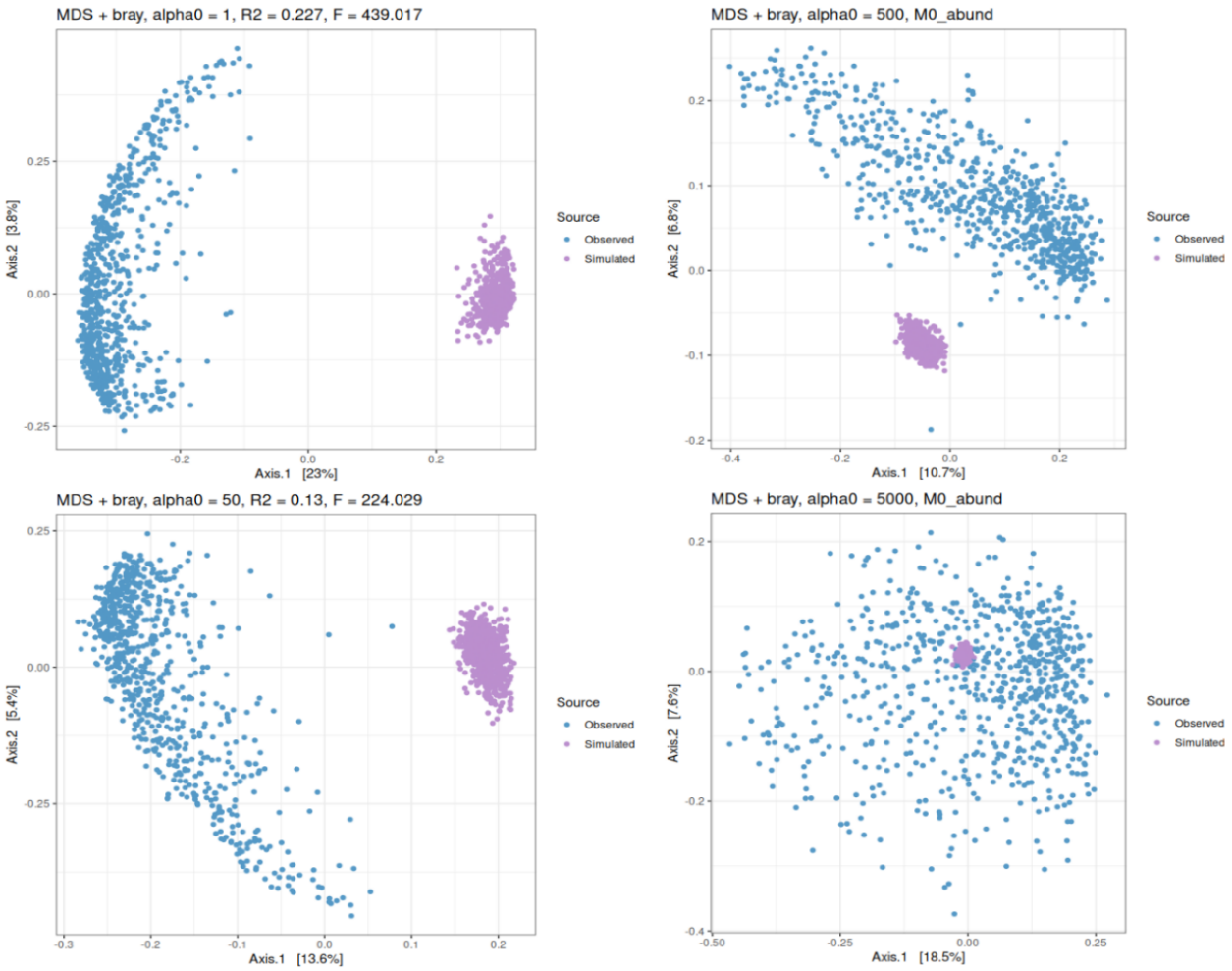


FIGURE 20 : MDS-Bray Curtis des microbiotes dirichlets (“simulated”) et G0 (“observed”) en fonction du choix d’$\alpha 0$ paramètre de dispersion.

E.5 Diversités α entre G0 et G1

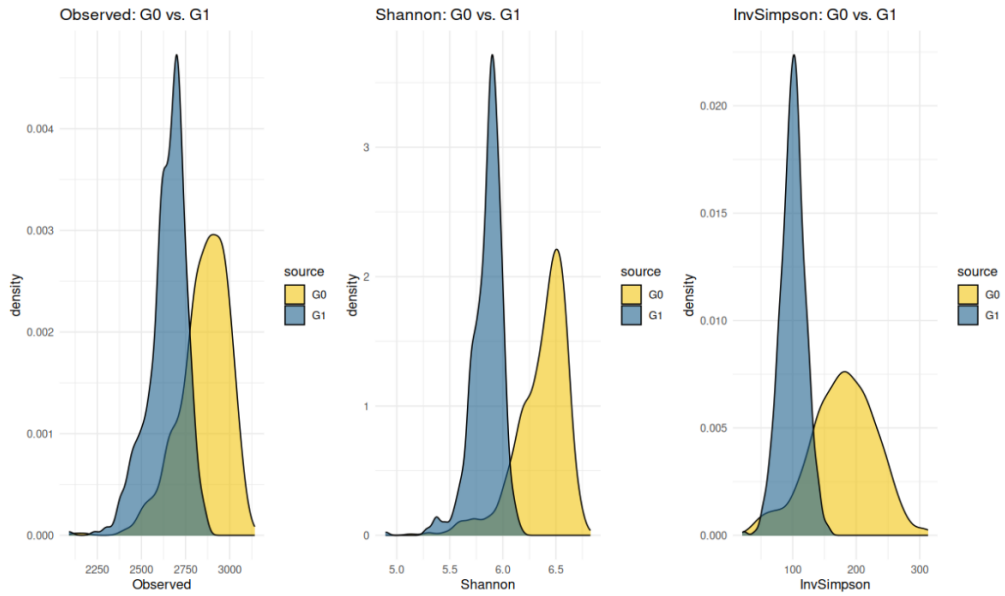


FIGURE 21 : Densité des 3 diversités calculées sur le microbiote G0 et G1. Analyse effectuée sur les données Bovines avec 750 individus et 4018 OTUs. Les paramètres de simulations par défaut ont été utilisés ormis `effect.size` et `size_rmultinom` calibrés sur les données.

E.6 Inégalité de Jensen

Avec $f :\rightarrow \mathbb{R}$ une fonction convexe. Pour tout $n \geq 2$, pour tous $x_1, ..., x_n \in I$, pour tous $\lambda_1, ..., \lambda_n \in [0, 1]$ avec $\sum_{i=1}^{n} \lambda_i = 1$, on a :

$$f(\sum_{i=1}^{n} \lambda_i x_i) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$$

Théorème : Soit X une variable aléatoire réelle intégrable. Soit ϕ une fonction convexe bornée inférieurement (i.e. telle que $\phi \geq a$ pour un certain réel a). Alors $\phi(\mathbb{E}(X)) \leq \mathbb{E}(\phi(X))$.

F Cas d'étude : condition contrôle

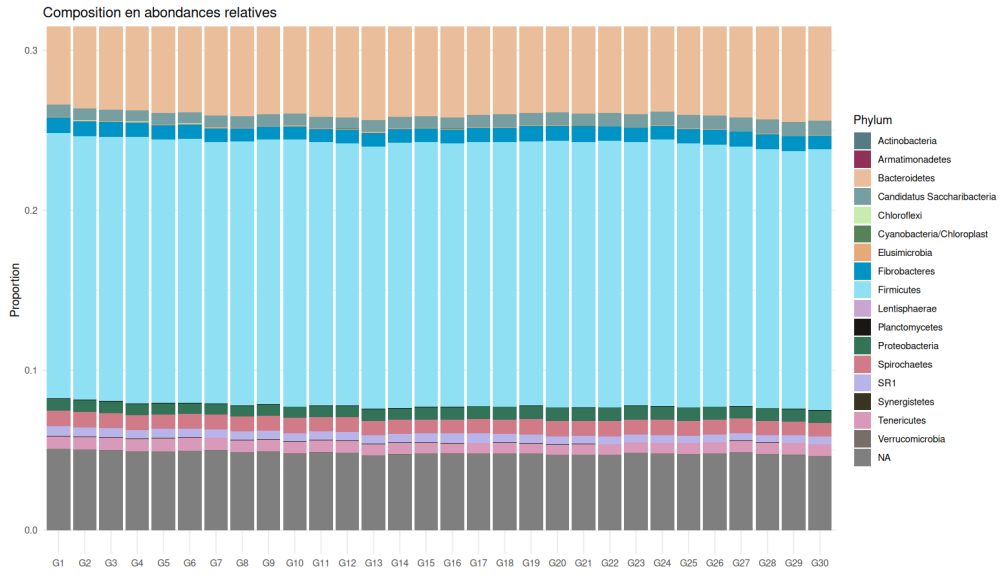


FIGURE 22 : Abondances relatives des microbiotes de 30 générations au niveau phylum, sans effets environnementaux. Analyse effectuée sur les données Bovines avec 750 individus et 3880 OTUS. Paramètre de simulation : aucuns OTUs sous contrôle génétique.

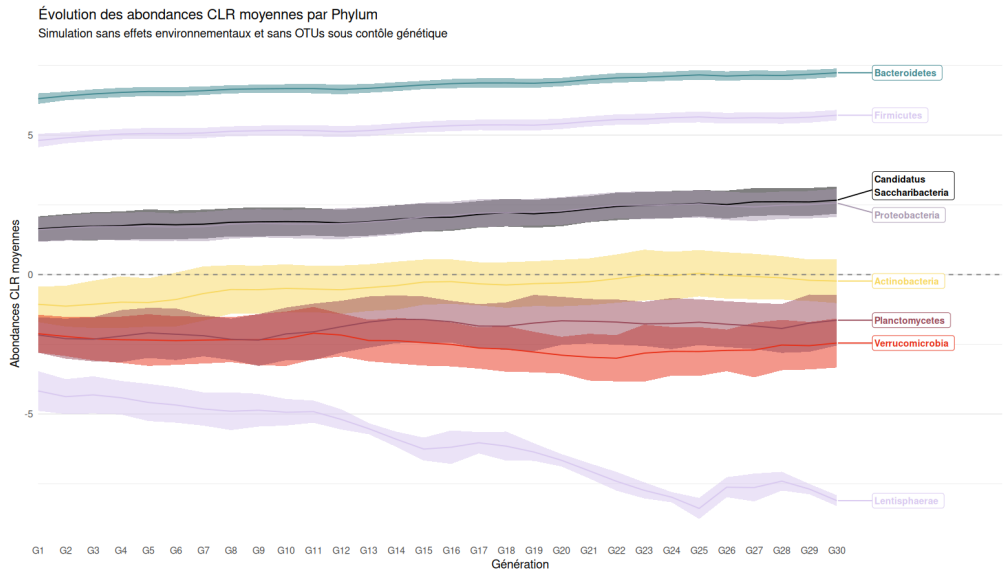


FIGURE 23 : Abondances CLR-transformées des microbiotes de 30 générations au niveau phylum, sans effets environnementaux. Analyse effectuée sur les données Bovines avec 750 individus et 3880 OTUS. Paramètre de simulation : aucuns OTUs sous contrôle génétique.

G Environnement R

```
 Session info
setting  value
version  R version 4.5.1 (2025-06-13)
os       Ubuntu 24.04.2 LTS
system   x86_64, linux-gnu
ui       RStudio
language (EN)
collate  fr_FR.UTF-8
ctype    fr_FR.UTF-8
tz       Europe/Paris
date     2025-08-29
rstudio  2024.12.1+563 Kousa Dogwood (desktop)
pandoc   3.1.3 @ /usr/bin/pandoc
quarto   1.7.33 @ /usr/local/bin/quarto

 Packages
package          * version  date (UTC) lib source
ade4               1.7-23   2025-02-14 [1] CRAN (R 4.5.1)
ape                5.8-1    2024-12-16 [1] CRAN (R 4.5.1)
attachment         0.4.5    2025-03-14 [1] CRAN (R 4.5.1)
bayesm             3.1-6    2023-09-23 [1] CRAN (R 4.5.1)
bigassertr         0.1.7    2025-06-27 [1] CRAN (R 4.5.1)
bigparallelr       0.3.2    2021-10-02 [1] CRAN (R 4.5.1)
bigsnpr          * 1.12.21  2025-08-21 [1] CRAN (R 4.5.1)
bigsparser         0.7.3    2024-09-06 [1] CRAN (R 4.5.1)
bigstatsr        * 1.6.2    2025-07-29 [1] CRAN (R 4.5.1)
Biobase            2.68.0   2025-04-15 [1] Bioconductor 3.21 (R 4.5.1)
BiocGenerics       0.54.0   2025-04-15 [1] Bioconductor 3.21 (R 4.5.1)
biomformat         1.36.0   2025-04-15 [1] Bioconductor 3.21 (R 4.5.1)
Biostrings         2.76.0   2025-04-15 [1] Bioconductor 3.21 (R 4.5.1)
cachem             1.1.0    2024-05-16 [2] CRAN (R 4.4.2)
circlize           0.4.16   2024-02-20 [1] CRAN (R 4.5.1)
cli                3.6.5    2025-04-23 [1] CRAN (R 4.5.1)
clue               0.3-66   2024-11-13 [1] CRAN (R 4.5.1)
cluster            2.1.8.1  2025-03-12 [4] CRAN (R 4.4.3)
codetools          0.2-20   2024-03-31 [4] CRAN (R 4.4.0)
colorspace         2.1-1    2024-07-26 [2] CRAN (R 4.4.2)
ComplexHeatmap   * 2.24.1   2025-06-25 [1] Bioconductor 3.21 (R 4.5.1)
compositions     * 2.0-8    2024-01-31 [1] CRAN (R 4.5.1)
cowplot            1.2.0    2025-07-07 [1] CRAN (R 4.5.1)
crayon             1.5.3    2024-06-20 [2] CRAN (R 4.4.2)
data.table       * 1.17.8   2025-07-10 [1] CRAN (R 4.5.1)
DEoptimR           1.1-3-1  2024-11-23 [1] CRAN (R 4.5.1)
desc               1.4.3    2023-12-10 [1] CRAN (R 4.5.1)
devtools         * 2.4.5    2022-10-11 [1] CRAN (R 4.5.1)
digest             0.6.37   2024-08-19 [2] CRAN (R 4.4.2)
dirmult            0.1.3-5  2022-03-21 [1] CRAN (R 4.5.1)
doParallel         1.0.17   2022-02-07 [1] CRAN (R 4.5.1)
doRNG              1.8.6.2  2025-04-02 [1] CRAN (R 4.5.1)
dplyr            * 1.1.4    2023-11-17 [2] CRAN (R 4.4.2)
ellipsis           0.3.2    2021-04-29 [1] CRAN (R 4.5.1)
evaluate           1.0.3    2025-01-10 [2] CRAN (R 4.4.2)
farver             2.1.2    2024-05-13 [2] CRAN (R 4.4.2)
fastmap            1.2.0    2024-05-15 [2] CRAN (R 4.4.2)
flock              0.7      2016-11-12 [1] CRAN (R 4.5.1)
forcats          * 1.0.0    2023-01-29 [2] CRAN (R 4.4.2)
foreach            1.5.2    2022-02-02 [1] CRAN (R 4.5.1)
fs                 1.6.5    2024-10-30 [2] CRAN (R 4.4.2)
fusen            * 0.7.1    2025-01-26 [1] CRAN (R 4.5.1)
generics           0.1.4    2025-05-09 [1] CRAN (R 4.5.1)
GenomeInfoDb       1.44.0   2025-04-15 [1] Bioconductor 3.21 (R 4.5.1)
GenomeInfoDbData   1.2.14   2025-07-16 [1] Bioconductor
GetoptLong         1.0.5    2020-12-15 [1] CRAN (R 4.5.1)
ggplot2          * 3.5.2    2025-04-09 [1] CRAN (R 4.5.1)
ggrepel          * 0.9.6    2024-09-07 [1] CRAN (R 4.5.1)
ggridges         * 0.5.6    2024-01-23 [1] CRAN (R 4.5.1)
ggtext           * 0.1.2    2022-09-16 [1] CRAN (R 4.5.1)
GlobalOptions      0.1.2    2020-06-10 [1] CRAN (R 4.5.1)
glue             * 1.8.0    2024-09-30 [2] CRAN (R 4.4.2)
gridtext           0.1.5    2022-09-16 [1] CRAN (R 4.5.1)
gtable             0.3.6    2024-10-25 [2] CRAN (R 4.4.2)
hms                1.1.3    2023-03-21 [2] CRAN (R 4.4.2)
htmltools          0.5.8.1  2024-04-04 [2] CRAN (R 4.4.2)
htmlwidgets        1.6.4    2023-12-06 [1] CRAN (R 4.5.1)
httpuv             1.6.15   2024-03-26 [2] CRAN (R 4.4.2)
httr               1.4.7    2023-08-15 [2] CRAN (R 4.4.2)
igraph             2.1.4    2025-01-23 [1] CRAN (R 4.5.1)
IRanges            2.42.0   2025-04-15 [1] Bioconductor 3.21 (R 4.5.1)
iterators          1.0.14   2022-02-05 [1] CRAN (R 4.5.1)
jsonlite           2.0.0    2025-03-27 [1] CRAN (R 4.5.1)
knitr              1.49     2024-11-08 [2] CRAN (R 4.4.2)
later              1.4.1    2024-11-27 [2] CRAN (R 4.4.2)
lattice            0.22-5   2023-10-24 [4] CRAN (R 4.3.3)
lazyeval           0.2.2    2019-03-15 [1] CRAN (R 4.5.1)
legendry         * 0.2.2    2025-05-30 [1] CRAN (R 4.5.1)
lifecycle          1.0.4    2023-11-07 [2] CRAN (R 4.4.2)
lubridate        * 1.9.4    2024-12-08 [2] CRAN (R 4.4.2)
```

```
magrittr        * 2.0.3    2022-03-30 [2] CRAN (R 4.4.2)
MASS              7.3-65   2025-02-28 [4] CRAN (R 4.4.3)
Matrix            1.7-3    2025-03-11 [4] CRAN (R 4.4.3)
matrixStats       1.5.0    2025-01-07 [1] CRAN (R 4.5.1)
memoise           2.0.1    2021-11-26 [2] CRAN (R 4.4.2)
mgcv              1.9-1    2023-12-21 [4] CRAN (R 4.3.2)
mime              0.13     2025-03-17 [1] CRAN (R 4.5.1)
miniUI            0.1.2    2025-04-17 [1] CRAN (R 4.5.1)
MoBPS           * 1.10.49  2025-07-16 [1] url (https://github.com/tpook92/MoBPS/raw/master/Previous%20versions/MoBPS_1.10.49.tar.gz)
multtest          2.64.0   2025-04-15 [1] Bioconductor 3.21 (R 4.5.1)
nlme              3.1-168  2025-03-31 [4] CRAN (R 4.4.3)
paletteer       * 1.6.0    2024-01-21 [1] CRAN (R 4.5.1)
patchwork       * 1.3.1    2025-06-21 [1] CRAN (R 4.5.1)
permute         * 0.9-8    2025-06-25 [1] CRAN (R 4.5.1)
phyloseq        * 1.52.0   2025-04-15 [1] Bioconductor 3.21 (R 4.5.1)
pillar            1.11.0   2025-07-04 [1] CRAN (R 4.5.1)
pkgbuild          1.4.8    2025-05-26 [1] CRAN (R 4.5.1)
pkgconfig         2.0.3    2019-09-22 [2] CRAN (R 4.4.2)
pkgdown         * 2.1.3    2025-05-25 [1] CRAN (R 4.5.1)
pkgload           1.4.0    2024-06-28 [1] CRAN (R 4.5.1)
plotly          * 4.11.0   2025-06-19 [1] CRAN (R 4.5.1)
plyr              1.8.9    2023-10-02 [1] CRAN (R 4.5.1)
png               0.1-8    2022-11-29 [1] CRAN (R 4.5.1)
profvis           0.4.0    2024-09-20 [1] CRAN (R 4.5.1)
promises          1.3.2    2024-11-28 [2] CRAN (R 4.4.2)
purrr           * 1.1.0    2025-07-10 [1] CRAN (R 4.5.1)
R6                2.6.1    2025-02-15 [1] CRAN (R 4.5.1)
RColorBrewer      1.1-3    2022-04-03 [2] CRAN (R 4.4.2)
Rcpp              1.1.0    2025-07-02 [1] CRAN (R 4.5.1)
readr           * 2.1.5    2024-01-10 [2] CRAN (R 4.4.2)
rematch2          2.1.2    2020-05-01 [2] CRAN (R 4.4.2)
remotes           2.5.0    2024-03-17 [1] CRAN (R 4.5.1)
reshape2          1.4.4    2020-04-09 [1] CRAN (R 4.5.1)
rhdf5             2.52.1   2025-06-08 [1] Bioconductor 3.21 (R 4.5.1)
rhdf5filters      1.20.0   2025-04-15 [1] Bioconductor 3.21 (R 4.5.1)
Rhdf5lib          1.30.0   2025-04-15 [1] Bioconductor 3.21 (R 4.5.1)
RITHMS          * 0.0.2    2025-07-16 [1] Github (SolenePety/RITHMS@549e4e6)
rjson             0.2.23   2024-09-16 [1] CRAN (R 4.5.1)
rlang             1.1.6    2025-04-11 [1] CRAN (R 4.5.1)
rmio              0.4.0    2022-02-17 [1] CRAN (R 4.5.1)
rngtools          1.5.2    2021-09-20 [1] CRAN (R 4.5.1)
robustbase        0.99-4-1 2024-09-27 [1] CRAN (R 4.5.1)
roxygen2        * 7.3.2    2024-06-28 [1] CRAN (R 4.5.1)
rstudioapi        0.17.1   2024-10-22 [2] CRAN (R 4.4.2)
S4Vectors         0.46.0   2025-04-15 [1] Bioconductor 3.21 (R 4.5.1)
scales          * 1.4.0    2025-04-24 [1] CRAN (R 4.5.1)
sessioninfo       1.2.3    2025-02-05 [1] CRAN (R 4.5.1)
shape             1.4.6.1  2024-02-23 [1] CRAN (R 4.5.1)
shiny             1.10.0   2024-12-14 [2] CRAN (R 4.4.2)
stringi           1.8.7    2025-03-27 [1] CRAN (R 4.5.1)
stringr         * 1.5.1    2023-11-14 [2] CRAN (R 4.4.2)
survival          3.8-3    2024-12-17 [4] CRAN (R 4.4.2)
tensorA           0.36.2.1 2023-12-13 [1] CRAN (R 4.5.1)
tibble          * 3.3.0    2025-06-08 [1] CRAN (R 4.5.1)
tidyr           * 1.3.1    2024-01-24 [2] CRAN (R 4.4.2)
tidyselect        1.2.1    2024-03-11 [2] CRAN (R 4.4.2)
tidyverse       * 2.0.0    2023-02-22 [2] CRAN (R 4.4.2)
timechange        0.3.0    2024-01-18 [2] CRAN (R 4.4.2)
tzdb              0.4.0    2023-05-12 [2] CRAN (R 4.4.2)
UCSC.utils        1.4.0    2025-04-15 [1] Bioconductor 3.21 (R 4.5.1)
urlchecker        1.0.1    2021-11-30 [1] CRAN (R 4.5.1)
usethis         * 3.1.0    2024-11-26 [1] CRAN (R 4.5.1)
vctrs             0.6.5    2023-12-01 [2] CRAN (R 4.4.2)
vegan           * 2.7-1    2025-06-05 [1] CRAN (R 4.5.1)
viridisLite       0.4.2    2023-05-02 [2] CRAN (R 4.4.2)
withr             3.0.2    2024-10-28 [2] CRAN (R 4.4.2)
xfun              0.52     2025-04-02 [1] CRAN (R 4.5.1)
xml2              1.3.6    2023-12-04 [2] CRAN (R 4.4.2)
xtable            1.8-4    2019-04-21 [2] CRAN (R 4.4.2)
XVector           0.48.0   2025-04-15 [1] Bioconductor 3.21 (R 4.5.1)
yaml              2.3.10   2024-07-26 [2] CRAN (R 4.4.2)
```

Glossaire

Analyse en Composantes Principales (ACP) : Méthode statistique de réduction de dimension des donénes tout en conservant la plus grande variance possible. Elle repose sur la décomposition en valeurs singulières (SVD) : les vecteurs propres définissent les axes principaux et les valeurs singulières quantifient la part de variance expliquée.

Centered Log-Ratio Transformation (CLR) : Transformation utilisée pour les données de composition (abondances relatives des microbiotes) qui permet de supprimer la contrainte de somme constante.

Diversité α : mesure de la richesse et de l'hétérogénéité des taxa au sein d'un même échantillon.

Diversité β : mesure des différences de composition microbienne entre plusieurs échantillons.

Héritabilité directe : Fraction de la variance phénotypique expliquée par la génétique de l'hôte. Effets génétiques directs sur le phénotype.

Héritabilité totale : Fraction de la variance phénotypique expliquée par les effets génétiques directs et les effets génétiques indirects (portion du microbiote ayant un impact sur le phénotype et étant sous contrôle génétique).

Holobionte : Un holobionte désigne l'ensemble constitué par un organisme multicellulaire et des microbes qu'il héberge.

Microbiabilité : fraction de la variance phénotypique expliquée par la composition du microbiote associé.

Multidimensional Scaling (MDS) : méthode de représentation des données basée sur les dissimilarités calculées entre individus. Une matrice de distances est construite puis recherche d'une dimension réduite tout en conservant au mieux les distances.

Operational Taxonomic Unit (OTU) : Les unités taxonomiques opérationnelles sont définies comme des groupes de séquences nucléotidiques très similaires, susceptibles de représenter un ou plusieurs organismes étroitement apparentés, généralement identifiés par un degré élevé d'identité nucléotidique (généralement > 97 %).

Prévalence (d'OTU) : Proportion d'échantillons dans lesquels une unité taxonomique opérationnelle (OTU) ou taxa est détecté.

Profondeur de séquençage : Nombre total de lectures (reads) obtenues pour un échantillon lors du séquençage. Permet de refléter la résolution et la sensibilité de la détection.

Quantitative Trait Locus (QTL) : Région du génome associé à la variation d'un caractère quantitatif.

Recrutement horizontal : acquisition du microbiote à partir de l'environnement ou des congénères.

Singular Value Decomposition (SVD) : Méthode permettant la factorisation d'une matrice en trois composantes : (1) une matrice de vecteurs propres (directions principales), (2) un ematrice diagonale des valeurs singulières (proportion de variance expliquée) et (3) une matrice de projection.

Single Nucleotide Polymorphism (SNP) : variation d'une seule paire de bases du génome (considérée comme marqueur génétique) entre individus.

Transmission verticale : transmission du microbiote de la mère à la descendance (gestion, mise bas, allaitement).

Bibliographie

Allaire JJ, Teague C, Scheidegger C, et al (2025) Quarto

Baniel A, Amato KR, Beehner JC, et al (2021) Seasonal shifts in the gut microbiome indicate plastic responses to diet in wild geladas. Microbiome 9 : https://doi.org/10.1186/s40168-020-00977-9

Calvin K, Dasgupta D, Krinner G, et al (2023) IPCC, 2023 : Climate Change 2023 : Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland. Intergovernmental Panel on Climate Change (IPCC)

Ceciliani F, Maggiolino A, Biscarini F, et al (2024) Heat stress has divergent effects on the milk microbiota of Holstein and Brown Swiss cows. Journal of Dairy Science 107:11639-11654. https://doi.org/10.3168/jds.2024-24976

Committee on Physiological Effects of Environmental Factors on Animals NRC (U.S.). (1971) A guide to environmental research on animals. National Academy of Sciences, Washington

Correia Sales GF, Carvalho BF, Schwan RF, et al (2021) Heat stress influence the microbiota and organic acids concentration in beef cattle rumen. Journal of Thermal Biology 97:102897. https://doi.org/10.1016/j.jtherbio.2021.102897

Cortes-Macías E, Selma-Royo M, García-Mantrana I, et al (2021) Maternal Diet Shapes the Breast Milk Microbiota Composition and Diversity : Impact of Mode of Delivery and Antibiotic Exposure. The Journal of Nutrition 151:330-340. https://doi.org/10.1093/jn/nxaa310

Cristóbal E, Ayuso SV, Justel A, Toro M (2013) Robust optima and tolerance ranges of biological indicators : a new method to identify sentinels of global warming. Ecological Research 29:55-68. https://doi.org/10.1007/s11284-013-1099-9

Déru V, Bouquet A, Hassenfratz C, et al (2020) Impact of a high-fibre diet on genetic parameters of production traits in growing pigs. Animal 14:2236-2245. https://doi.org/10.1017/s1751731120001275

Déru V, Bouquet A, Zemb O, et al (2022) Genetic relationships between efficiency traits and gut microbiota traits in growing pigs being fed with a conventional or a high-fiber diet. Journal of Animal Science 100 : https://doi.org/10.1093/jas/skac183

Difford GF, Plichta DR, Løvendahl P, et al (2018) Host genetics and the rumen microbiome jointly associate with methane emissions in dairy cows. PLOS Genetics 14:e1007580. https://doi.org/10.1371/journal.pgen.1007580

Digital Scholarship C for (2025) Zotero : Reference Management Software

Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ (2017) Microbiome Datasets Are Compositional : And This Is Not Optional. Frontiers in Microbiology 8 : https://doi.org/10.3389/fmicb.2017.02224

Kassambara A, Mundt F (2016) factoextra : Extract and Visualize the Results of Multivariate Data Analyses. CRAN : Contributed Packages

Kim SH, Ramos SC, Valencia RA, et al (2022) Heat Stress : Effects on Rumen Microbes and Host Physiology, and Strategies to Alleviate the Negative Impacts on Lactating Dairy Cows. Frontiers in Microbiology 13 : https://doi.org/10.3389/fmicb.2022.804562

Koch F, Otten W, Sauerwein H, et al (2023) Mild heat stress–induced adaptive immune response in blood mononuclear cells and leukocytes from mesenteric lymph nodes of primiparous lactating Holstein cows. Journal of Dairy Science 106:3008-3022. https://doi.org/10.3168/jds.2022-22520

Koch F, Reyer H, Görs S, et al (2024) Heat stress and feeding effects on the mucosa-associated and digesta microbiome and their relationship to plasma and digesta fluid metabolites in the jejunum of dairy cows. Journal of Dairy Science 107:5162-5177. https://doi.org/10.3168/jds.2023-24242

Landi V, Maggiolino A, Hidalgo J, et al (2024) Effect of transgenerational environmental condition on genetics parameters of Italian Brown Swiss. Journal of Dairy Science 107:1549-1560. https://doi.org/10.3168/jds.2023-23741

Lê S, Josse J, Husson F (2008) FactoMineR : AnRPackage for Multivariate Analysis. Journal of Statistical Software 25 : https://doi.org/10.18637/jss.v025.i01

Li H, Li R, Chen H, et al (2019) Effect of different seasons (spring vs summer) on the microbiota diversity in the feces of dairy cows. International Journal of Biometeorology 64:345-354. https://doi.org/10.1007/s00484-019-01812-z

Mader TL, Davis MS, Brown-Brandl T (2006) Environmental factors influencing heat stress in feedlot cattle1, 2. Journal of Animal Science 84:712-719. https://doi.org/10.2527/2006.843712x

Maurice CF, Knowles SCL, Ladau J, et al (2015) Marked seasonal variation in the wild mouse gut microbiota. The ISME Journal 9:2423-2434. https://doi.org/10.1038/ismej.2015.53

McMurdie PJ, Holmes S (2013) phyloseq : An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE 8:e61217. https://doi.org/10.1371/journal.pone.0061217

Oksanen J, Simpson GL, Blanchet FG, et al (2025) vegan : Community Ecology Package

Pérez-Enciso M, Zingaretti LM, Ramayo-Caldas Y, Campos G de los (2021) Opportunities and limits

of combining microbiome and genome data for complex trait prediction. Genetics Selection Evolution 53 : https://doi.org/10.1186/s12711-021-00658-7

Pety S, Mariadassou M, David I, Rau A (2025) RITHMS : An advanced stochastic framework for the simulation of transgenerational hologenomic data

Pook T, Schlather M, Simianer H (2020) MoBPS - Modular Breeding Program Simulator. G3 Genes|Genomes|Genetics 10:1915-1918. https://doi.org/10.1534/g3.120.401193

Privé F, Aschard H, Ziyatdinov A, Blum MGB (2018) Efficient analysis of large-scale genome-wide data with two R packages : bigstatsr and bigsnpr. Bioinformatics 34:2781-2787. https://doi.org/10.1093/bioinformatics/bty185

R Core Team (2025) The Comprehensive R Archive Network

Rhoads RP, La Noce AJ, Wheelock JB, Baumgard LH (2011) Short communication : Alterations in expression of gluconeogenic genes during heat stress and exogenous bovine somatotropin administration. Journal of Dairy Science 94:1917-1921. https://doi.org/10.3168/jds.2010-3722

Rochette S, Guyader V, Mansiaux Y (2025) fusen : Build a Package from Rmarkdown Files

Rutayisire E, Huang K, Liu Y, Tao F (2016) The mode of delivery affects the diversity and colonization pattern of the gut microbiota during the first year of infants' life : a systematic review. BMC Gastroenterology 16 : https://doi.org/10.1186/s12876-016-0498-0

Shterzer N, Rothschild N, Sbehat Y, et al (2023) Vertical transmission of gut bacteria in commercial chickens is limited. Animal Microbiome 5 : https://doi.org/10.1186/s42523-023-00272-6

Souza VC, Moraes LE, Baumgard LH, et al (2023) Modeling the effects of heat stress in animal performance and enteric methane emissions in lactating dairy cows. Journal of Dairy Science 106:4725-4737. https://doi.org/10.3168/jds.2022-22658

Tao S, Orellana Rivas RM, Marins TN, et al (2020) Impact of heat stress on lactational performance of dairy cows. Theriogenology 150:437-444. https://doi.org/10.1016/j.theriogenology.2020.02.048

Theis KR, Dheilly NM, Klassen JL, et al (2016) Getting the Hologenome Concept Right : an Eco-Evolutionary Framework for Hosts and Their Microbiomes. mSystems 1 : https://doi.org/10.1128/msystems.00028-16

Verma KK, Song X, Kumari A, et al (2024) Climate change adaptation : Challenges for agricultural sustainability. Plant, Cell & Environment 48:2522-2533. https://doi.org/10.1111/pce.15078

Wallace RJ, Sasson G, Garnsworthy PC, et al (2019) A heritable subset of the core rumen microbiome dictates dairy cow productivity and emissions. Science Advances 5 : https://doi.org/10.1126/sciadv.aav8391

Wheelock JB, Rhoads RP, VanBaale MJ, et al (2010) Effects of heat stress on energetic metabolism in lactating Holstein cows. Journal of Dairy Science 93:644-655. https://doi.org/10.3168/jds.2009-2295

Wickham H, Bryan J (2023) R Packages, 2nd edition. O'Reilly Media

Wickham H, Bryan J, Barrett M, Teucher A (2025a) usethis : Automate Package and Project Setup

Wickham H, Danenberg P, Csárdi G, Eugster M (2024) roxygen2 : In-Line Documentation for R

Wickham H, Hesselberth J, Salmon M, et al (2025b) pkgdown : Make Static HTML Documentation for a Package

Wickham H, Hester J, Chang W, Bryan J (2022) devtools : Tools to Make Developing R Packages Easier

Williams CE, Williams CL, Logan ML (2023) Climate change is not just global warming : Multi-dimensional impacts on animal gut microbiota. Microbial Biotechnology 16:1736-1744. https://doi.org/10.1111/1751-7915.14276

Zang X-W, Sun H-Z, Xue M-Y, et al (2022) Heritable and Nonheritable Rumen Bacteria Are Associated with Different Characters of Lactation Performance of Dairy Cows. mSystems 7 : https://doi.org/10.1128/msystems.00422-22

Zilber-Rosenberg I, Rosenberg E (2008) Role of microorganisms in the evolution of animals and plants : the hologenome theory of evolution. FEMS Microbiology Reviews 32:723-735. https://doi.org/10.1111/j.1574-6976.2008.00123.x