Étude du rôle fonctionnel du microbiote intestinal dans l'adaptation à un régime sous optimal et dans l'efficience alimentaire de la poule pondeuse en utilisant une approche multi-omique

Maria Bernard

THESE DE DOCTORAT

NNT : 2024UPASB075

université PARIS-SACLAY

Étude du rôle fonctionnel du microbiote intestinal dans l'adaptation à un régime sous optimal et dans l'efficience alimentaire de la poule pondeuse en utilisant une approche multi-omique.

Study of the functional role of the gut microbiota in adaptation to a sub-optimal diet and in feed efficiency in laying hens using a multi-omics approach.

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 581 agriculture, alimentation, biologie, environnement et santé (ABIES)
Spécialité de doctorat : Génétique Animale
Graduate School : Biosphera. Référent : AgroParisTech

Thèse préparée dans l'UMR **GABI** (Université Paris-Saclay, INRAE, AgroParisTech), et dans l'UMR **GenPhySE** (Université de Toulouse 3, INRAE, ENVT) sous la direction de **Tatiana ZERJAL**, Chargée de recherche (ADR), le co-encadrement de **Géraldine PASCAL**, Ingénieure de recherche, et le co-encadrement de **Fanny CALENGE**, Chargée de recherche

Thèse soutenue à Paris-Saclay, le 11 décembre 2024, par

Maria BERNARD

Composition du Jury

Membres du jury avec voix délibérative

Catherine LARZUL Directrice de recherche, INRAE (Université de Toulouse 3)	Présidente
Núria MACH Chargée de recherche (HDR), INRAE (Université de Toulouse 3)	Rapporteur & Examinatrice
Pierre PEYRET Professeur, Université Clermont Auvergne	Rapporteur & Examinateur
Florent KEMPF Ingénieur de recherche, INRAE (Université de Tours)	Examinateur
Sophie SCHBATH Directrice de recherche, INRAE (Université Paris-Saclay)	Examinatrice

université PARIS-SACLAY
ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)

Titre : Etude du rôle fonctionnel du microbiote intestinal dans l'adaptation à un régime sous optimal et dans l'efficience alimentaire de la poule pondeuse en utilisant une approche multi-omique.

Mots clés : Microbiote intestinal, poule pondeuse, efficience alimentaire, régime sous-optimal, analyse multi-omique

Résumé : La poule est l'espèce animale d'élevage la plus exploitée dans le monde, et les œufs, en plus d'avoir des qualités nutritives indéniables, représentent, pour l'homme, la ressource d'origine animale la moins coûteuse. Les industries avicoles s'efforcent d'optimiser leur production, notamment en améliorant l'efficience alimentaire, un caractère important pour assurer la rentabilité des élevages tout en réduisant leur impact environnemental. Ce caractère complexe est sous l'influence de la génétique et de l'environnement, mais de plus en plus d'études confirment également le rôle du microbiote intestinal. Celui-ci participe à la dégradation des aliments et produit de nombreux métabolites qui protègent l'hôte et influencent son métabolisme. Ainsi, plusieurs associations ont été établies entre le microbiote et des caractères de croissance, de santé, de performance et d'efficience alimentaire.

Cette thèse propose une analyse du microbiote cæcal de poules pondeuses issues de deux lignées sélectionnées de façon divergente sur leur efficience alimentaire. Le régime alimentaire influençant à la fois l'efficience alimentaire et la composition du microbiote, ces deux lignées ont été nourries avec deux régimes alimentaires, l'un optimal et l'autre réduit en énergie et enrichi en fibres. Pour caractériser les écosystèmes microbiens, trois approches omiques ont été utilisées : le séquençage métabarcoding ciblant le gène de l'ARNr 16S, la métagénomique (séquençage complet des ADN) et la métatranscriptomique (séquençage complet des ARN). Alors que le métabarcoding permet une caractérisation des communautés microbiennes et des fonctions moins résolutive que les deux autres, ces dernières sont plus coûteuses et présentent des défis expérimentaux et méthodologiques importants.

Cette thèse a donc eu comme objectifs de répondre à des questions biologiques et méthodologiques : i) identifier le rôle du microbiote cæcal dans l'efficience alimentaire des poules pondeuses, ii) évaluer l'impact du régime alimentaire sur le microbiote et sur son rôle dans l'efficience alimentaire, iii) comparer les différentes approches omiques pour répondre à ces questions.

L'analyse de ces données a permis de mettre en évidence des différences de compositions taxonomiques mais également fonctionnelles, à la fois selon la lignée et selon le régime. En outre, les hypothèses du rôle du microbiote dans l'efficience alimentaire des animaux sont conditionnées au régime utilisé pour nourrir les animaux. Elles impliquent notamment des capacités de dégradation de carbohydrates variés (amidon ou fibres indigestes) et aboutissent à une production différenciée d'acides gras à chaîne courte connus pour influencer le métabolisme de l'hôte. Du point de vue de la méthodologie d'analyse, des outils et des stratégies ont été comparés pour mettre en place une chaîne de traitement de ces séquences, tout en mettant en évidence des verrous et des difficultés méthodologiques, nécessitant de futurs développements.

Cette thèse apporte de nouveaux éléments sur le rôle du microbiote dans l'efficience alimentaire des poules pondeuses, avec un accent particulier sur ses fonctions métaboliques. Elle souligne également les avantages et les limites des trois techniques omiques utilisées. Des analyses complémentaires sont toutefois nécessaires pour intégrer de façon plus complète les résultats issus de ces différentes approches omiques et pour affiner l'identification des activités microbiennes impliquées.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)

Title : Study of the functional role of the gut microbiota in adaptation to a sub-optimal diet and in feed efficiency in laying hens using a multi-omics approach.

Keywords : Gut microbiota, laying hens, feed efficiency, sub-optimal diet, multi-omics analysis

Abstract : Chickens are the most widely exploited farmed animal species in the world, and eggs, in addition to their undeniable nutritional qualities, represent the least expensive animal-based resource for human consumption. The poultry industry strives to optimize production, in particular by improving feed efficiency, an important factor in ensuring farm profitability while reducing environmental impact. This complex trait is influenced by genetics and the environment, but more and more studies are also confirming the role of the intestinal microbiota. It is involved in the breakdown of food and produces numerous metabolites that protect the host and influence its metabolism. Several associations have been established between the microbiota and growth, health, performance and feed efficiency traits.

This thesis presents an analysis of the caecal microbiota of laying hens from two lines divergently selected for feed efficiency. As diet influences both feed efficiency and microbiota composition, these two lines were fed two diets, one optimal and the other reduced in energy and enriched in fiber. To characterize the microbial ecosystems, three omics approaches were used: metabarcoding sequencing targeting the 16S rRNA gene, metagenomics (full DNA sequencing) and metatranscriptomics (full RNA sequencing). While metabarcoding enables less resolutive characterization of microbial communities and functions than the other two, the latter are more costly and present significant experimental and methodological challenges.

This thesis therefore aims to answer biological and methodological questions: i) to identify the role of the caecal microbiota in the feed efficiency of laying hens, ii) to assess the impact of diet on the microbiota and its role in feed efficiency, iii) to compare different omics approaches to answer these questions.

Analysis of these data revealed differences in taxonomic and functional compositions, both by line and by diet. Furthermore, hypotheses concerning the role of the microbiota in animal feed efficiency are conditioned by the diet used to feed the animals. In particular, they involve the capacity to degrade various carbohydrates (starch or indigestible fibers) and result in the differentiated production of short-chain fatty acids known to influence host metabolism. From a methodology analysis perspective, tools and strategies have been compared to establish a processing chain for these sequences, while highlighting obstacles and difficulties requiring future development.

This thesis provides new insights into the role of the microbiota in the feed efficiency of laying hens, with particular emphasis on its metabolic functions. It also highlights the advantages and limitations of the three omics techniques used. Further analysis is needed, however, to integrate the results from these different omics approaches more completely, and to refine the identification of the microbial activities involved.

REMERCIEMENTS

Voilà, nous y sommes, un peu plus de trois années après avoir pris la décision, un peu folle, de réaliser cette thèse. Trois années durant lesquelles vous avez été nombreux à m'encourager, me soutenir, me conseiller aussi, ou me suggérer de nouvelles idées. Trois années durant lesquelles vous m'avez permis de me consacrer à ce projet, vous m'avez fait évoluer et c'était un de mes grands objectifs personnels. En espérant que ce travail sera à la hauteur de toutes ces interactions professionnelles ou amicales, il est temps maintenant de vous dire officiellement, mais surtout très sincèrement, Merci.

Pour commencer, j'aimerais remercier les membres de mon jury, qui ont accepté d'évaluer ce travail : **Núria MACH** et **Pierre PEYRET** en qualité de rapporteurs, **Florent KEMPF, Catherine LARZUL** et **Sophie SCHBATH** en qualité d'examinateurs.

Merci également aux membres de mon comité de thèse : **Sylvie Combes**, **Mahendra Mariadassou**, **Jordi Estellé** et **Xavier Rognon**. Ces comités ont représenté de véritables jalons pour poser les réalisations et avancer de nouvelles idées. Merci pour le temps que vous m'avez consacré pendant ces comités, mais également en dehors.

Je pense également aux nombreux échanges ponctuels que j'ai pu avoir avec différentes personnes d'unités INRAE variées, mais également en dehors : GetPlaGe, GenoToul-bioinfo, ProbiHôte, Metagenopolis, NED, et le Génoscope. Merci pour ces échanges riches qui m'ont permis d'avancer sur une multiplicité de points techniques.

Un très grand Merci, à l'ensemble de l'**unité GABI,** des **équipes SIGENAE** et **GIBBS** et du groupe **FROGS**. Un merci officiel d'abord, car j'ai eu la chance d'être soutenue sans réserve par la direction de l'unité et de mes équipes de rattachement pour réaliser cette thèse. Et un merci ou plutôt de nombreux mercis plus amicaux à chacun individuellement. Vous êtes trop nombreux pour que je puisse vous nommer un par un et quelque part cela m'arrange (je ne voudrais oublier personne). Cela fait plus de 12 ans que j'ai intégré l'unité et je suis passée par trois bâtiments, et quatre équipes. Cela illustre sans doute les nombreux messages que vous m'avez adressés à l'occasion d'évènements professionnels, mais aussi volontairement pour simplement prendre des nouvelles. Je crois que je n'ai pas toujours été à la hauteur de ces petits messages, mais sachez, toutes et tous, que cela m'a beaucoup touchée (… vraiment beaucoup !), cela m'a redonné des forces parfois et de la volonté pour continuer. Un chaleureux merci à chacun.

Permettez-moi toutefois de m'adresser à certains d'entre vous tout particulièrement, mais n'y voyez aucun odre d'importance dans l'ordre d'apparition.

Mathieu, co-bureau depuis longtemps, tu as été embarqué un peu malgré toi dans cette aventure en prenant la suite de mes projets … presque terminés (hic!). Tu as eu le plaisir de me voir passer par toutes les émotions (!!) et chaque fois tu m'as redonné le petit coup de pouce pour rebondir, et tu as partagé chacune de mes satisfactions avec sincérité. Merci beaucoup pour ton soutien. De même, la mare aux grenouilles a pris le petit (?) surplus de travail que je leur ai gentiment laissé. Merci à **Olivier, Lucas, Vincent**, **Gabryelle** et **Géraldine** pour nos discussions pleines d'humour qui font relâcher la pression, nos échanges scientifiques, et surtout notre amitié qui me font avancer.

A toutes l'équipe GIBBS et en particulier à **Andrea**, **Gwendal**, **Denis**, **Alicia**, **Quentin**, notre diversité de compétences a été une source incroyable de discussions, et de support, le tout avec une sacrée dose d'humour dont nous cultivons le style (!!). A vous tous, merci beaucoup.

A **Nicolas** et **Jean-Luc** qui m'ont permis de remettre une blouse blanche, mais pas de manipuler une pipette (chacun son métier !), et c'est très bien comme ça. J'ai vraiment apprécié me plonger (… en surface) dans le monde de l'expérimentation avec vous, bien que je me passerais volontiers de trier/peser encore une fois ces échantillons, et vous aussi, j'imagine. Merci à tous les deux.

Je remercie également **ma famille** et **mes amis**, qui ont su s'intéresser, m'écouter, m'encourager et me soutenir dans cette aventure, et qui m'ont permis de m'en échapper à l'occasion. Un merci tout particulier à **Tya**, **Charlie** et **Inès** qui, du haut de leur petite dizaine d'années, se sont également intéressées à la thèse de Tata, une touche rafraîchissante! A vous tous, je vous dis merci, merci et merci, j'ai hâte de vous retrouver.

Enfin, même si l'ordre n'avait pas d'importance jusque-là, je tenais à terminer en remerciant **Tatiana Zerjal**, **Géraldine Pascal** et **Fanny Calenge**, mes trois directrices de thèse. J'ai eu la chance que vous m'ayez accordé votre confiance et votre temps pour m'accompagner dans la réalisation de ces trois années de recherche. Vous avez formé le plus équilibré des trios, sur vos compétences scientifiques, et sur vos points de vue. Plus personnellement, Tatiana, une présence presque quotidienne, tu as été un soutien personnel incroyable que je n'oublierai pas. Nous avons appris à nous connaître, et j'espère que cette thèse n'est pas le dernier projet sur lequel nous nous retrouverons. Géraldine, c'était un petit défi de réaliser cela toutes les deux, alors que nous nous connaissons depuis de nombreuses années. Je crois que l'on peut dire que nous avons réussi, et je suis sincèrement ravie que tu aies fait partie de ce projet. Fanny, je retiens tout particulièrement de nos discussions ton optimisme dans la façon de présenter les choses, ce qui a eu le don de me faire avancer plus sereinement (en particulier pendant la rédaction). Un très grand merci à toutes les trois.

A Vivi, à Françis.

‘Si tout est significatif, plus rien ne l'est.’

Denis Laloë 2024

TABLE DES MATIERES

Table des figures 1

Table des Tableaux 4

A. Introduction 5

A.1. Contexte historique et agro-socio-économique de la poule pondeuse 5

A.1.1. Domestication et répartition mondiale de la poule domestique 5

A.1.2. Pratiques d'élevage et importance des œufs en nutrition humaine 6

A.1.3. Importance et évolution de la production d'œufs vis-à-vis des autres productions animales 8

A.2. Facteurs d'évolution de la production d'œufs 10

A.2.1. La sélection génétique : focus sur l'efficience alimentaire 10

A.2.2. Le régime alimentaire : un facteur clé de la production 13

A.2.3. Evolutions des performances et mise à jour des objectifs de recherche et développement 14

A.3. Le microbiote intestinal : généralités et focus sur la poule pondeuse 16

A.3.1. Microbiote : Définitions et interactions avec l'hôte 16

A.3.2. Les fonctions et productions du microbiote intestinal 18

A.3.3. L'appareil digestif de la poule 19

A.3.4. Composition du microbiote intestinal le long du tractus digestif de la poule 20

A.3.5. Principaux facteurs de variation du microbiote intestinal de poule 22

A.4. Evolution des méthodes d'analyses permettant d'explorer les écosystèmes microbiens 24

A.5. Méthodes contemporaines basées sur le séquençage 28

A.5.1. Le métabarcoding : séquençage d'un gène marqueur 29

A.5.2. La métagénomique : séquençage complet des génomes 38

A.5.3. La métatranscriptomique : séquençage complet des transcriptomes 43

B. Objectifs de la thèse 47

C. Démarche expérimentale et acquisition des données 50

C.1. Des lignées divergentes pour l'efficience alimentaire comme modèle expérimental 50

C.2. Mise en place de l'expérimentation animale et échantillonnage 51

C.3. Acquisition des données de séquences et processus d'analyses de données 53

C.3.1. Caractérisation du microbiote cæcal par séquençage ciblé de l'ARNr 16S 54

C.3.2. Caractérisation du microbiote cæcal par séquençage complet de l'ADN 57

C.3.3. Caractérisation du microbiote cæcal par séquençage complet de l'ARN 61

C.3.4. Les contrôles qualités préliminaires 64

D. Partie 1 : L'analyse de données de métabarcoding révèle des interactions hôtes-microbiotes dépendantes du régime ____ *70*

D.1. Contexte et objectifs ____ **71**

D.2. Matériels et méthodes ____ **72**

D.3. Travaux préliminaires à l'analyse d'inférence fonctionnelle ____ **73**

D.4. Résultats et Discussion ____ **75**

D.5. Valorisations scientifiques ____ **78**

E. Partie 2 : Constitution d'un catalogue enrichi de taxonomies et de fonctions ____ *96*

E.1. Contexte et objectifs ____ **96**

E.2. Mise au point méthodologique d'analyse de novo et d'intégration du catalogue de gènes et de MGS MetaChick ____ **100**

E.2.1. Comparaison de stratégies d'analyse de novo par assemblage individuel ou co-assemblage 100

E.2.2. Intégration du catalogue MetaChick et des analyses de novo de co-assemblage ____ 105

E.2.3. Métagénomique quantitative fonctionnelle et taxonomique : méthodologie et filtres ____ 107

E.3. Effet de la lignée et du régime alimentaire sur la composition microbienne et fonctionnelle : comparaison de la métagénomique et du métabarcoding ____ **112**

E.3.1. La métagénomique améliore la résolution taxonomique mais les affiliations taxonomiques divergent partiellement entre analyses omiques ____ 112

E.3.2. Les fonctions du microbiote sont globalement conservées malgré une diversité taxonomique variable ____ 118

E.4. Bilan des résultats et réflexion sur le développement méthodologique ____ **125**

F. Partie 3 : Analyse de données de métatranscriptomique : différenciation fonctionnelle des microbiotes cæcaux ____ *127*

F.1. Contexte et objectifs ____ **127**

F.2. La richesse et la diversité fonctionnelle à l'échelle des profils d'expression ____ **129**

F.3. Evaluation de la contribution des profils d'abondance dans les profils d'expression des fonctions ____ **134**

F.3.1. Comparaison des fonctions différentiellement abondantes et/ou exprimées en fonction de la lignée et du régime ____ 134

F.3.2. Analyse intégrative des profils d'abondances et d'expression ____ 137

F.4. Exploration des fonctions du microbiote cæcal différentiellement exprimées entre lignées et/ou régimes alimentaires _____ 140

F.4.1. Mise en évidence des voies métaboliques et catégories fonctionnelles les plus différenciées 140

F.4.2. Influence de caractéristiques physiques et génétiques de différents types de micro-organismes _____ 144

F.4.3. Le transport et le métabolisme des carbohydrates tendent à illustrer les différences d'activités du microbiote entre lignées et entre régimes _____ 147

F.5. Bilan des résultats et réflexion sur le développement méthodologique _____ 151

G. Productions, Valorisations, Perspectives _____ 154

G.1. De l'utilisation des outils bioinformatiques au développement de nouvelles fonctionnalités _____ 154

G.2. Finalisation et valorisation de l'analyse comparative des données omiques _____ 156

G.3. Tenir compte d'*a priori* pour l'analyse des métatranscriptomes _____ 158

H. Discussion _____ 161

H.1. Le rôle du microbiote cæcal dans l'efficience alimentaire des poules pondeuses : origine de sa différenciation _____ 163

H.2. Importance de l'impact du régime : défis et opportunités _____ 165

H.3. Vers une description exhaustive des communautés, des fonctions et de leur association _____ 167

H.3.1. Focus vers l'exhaustivité taxonomique _____ 167

H.3.2. De l'annotation fonctionnelle à l'identification taxonomique des contributeurs _____ 169

I. Conclusion et Bilan personnel _____ 172

Annexes _____ 175

J.1. Autres communications dans lesquelles je suis co-auteure _____ 175

J.2. Liste des figures annexes _____ 192

Bibliographie _____ 196

TABLE DES FIGURES

Figure A-1: Répartition des populations sauvages du genre Gallus. 5

Figure A-2: Répartition du nombre poules d'élevage en 2015 à travers le monde. 6

Figure A-3: Représentation de la production mondiale de viande et d'œufs. 9

Figure A-4 : Représentation schématique des mesures d'efficience alimentaire. 11

Figure A-5: Description du tractus digestif de la poule et principaux genres bactériens du microbiote intestinal. 20

Figure A-6: Abondances relatives des familles des quatre phyla majoritaires du microbiote intestinal de poule pondeuse. 22

Figure A-7 : Schéma des indices de diversité alpha et bêta. 26

Figure A-8 : Différents protocoles d'analyses du microbiote. 27

Figure A-9: Nombre de publications pour l'étude du microbiote intestinal chez l'homme ou la poule en fonction du protocole de séquençage utilisé. 28

Figure A-10: Schéma du protocole expérimental et analytique du métabarcoding. 30

Figure A-11: Représentation des biais naturels et techniques du métabarcoding et de solutions correctives. 36

Figure A-12 : Schéma de l'analyse *de novo* des séquences de métagénomique. 39

Figure A-13 : Schéma du protocole expérimental de métatranscriptomique. 45

Figure B-1: Facteurs d'influence du microbiote intestinal et de l'efficience alimentaire des poules. 47

Figure C-1: Evolution de la consommation résiduelle d'aliments (RFI) des lignées R+ et R-. 50

Figure C-2 : Représentation du plan d'expérimentation animale issu du projet ChickStress. 52

Figure C-3: Nombre d'animaux séquencés par type de données omiques. 53

Figure C-4 : Entropie et taille des régions hypervariables du gène de l'ARNr 16S. 55

Figure C-5: Représentation graphique des outils et paramètres de FROGS utilisés dans cette étude. 56

Figure C-6: Représentation graphique des outils et paramètres de metagWGS utilisés dans notre étude. 59

Figure C-7 : Proportion des lectures ribosomales, non ribosomales et alignées sur le catalogue MetaChick. 62

Figure C-8: Identification de potentiels échantillons aberrants et analyse de reproductibilité des échantillons re-séquencés. 65

Figure C-9 : Distribution du nombre de lectures et du taux d'alignement sur le catalogue MetaChick. 67

Figure D-1 : Distribution des valeurs de NSTI en fonction du pourcentage d'identité et de la couverture de l'alignement. 74

Figure E-1: Schéma du processus de traitement bioinformatique *de novo* des séquences issues de séquençage métagénomique avec intégration d'une référence externe. _____ 97
Figure E-2 : Schéma du processus de traitement de métagénomique quantitative. _____ 107
Figure E-3 : Distribution des abondances et richesse des MGS du catalogue MetaChick selon deux méthodes de quantification. _____ 109
Figure E-4 : Abondance relative des phyla dans chacun des groupes « lignées x régime ». _____ 113
Figure E-5 : Indices de richesse et de diversité des MGS selon la lignée et le régime._____ 114
Figure E-6 : Nombre de MGS différentiellement abondant et distribution des log2foldchange (LFC). _____ 117
Figure E-7 : Nombre de fonctions KEGG par catégories fonctionnelles._____ 120
Figure E-8 : Nombre de gènes et de familles de chaque classe de CAZymes, quantifiés sur nos échantillons cæcaux de poule. _____ 122
Figure E-9 : Richesse et diversité des fonctions KEGG et des CAZymes selon la lignée et le régime. 123
Figure E-10 : Effet du régime sur les distances de Bray Curtis calculées sur les MGS, les fonctions KEGG et les familles de CAZymes potentielles._____ 124
Figure F-1 : Proportion des lectures ribosomales, non ribosomales, provenant de l'hôte ou du microbiote cæcal. _____ 129
Figure F-2 : Courbes de raréfaction fonctionnelle sur les données de métatranscriptomique.____ 131
Figure F-3: Richesse et diversité des fonctions KEGG exprimées selon la lignée et le régime. ____ 132
Figure F-4 : Distribution des distances de Bray Curtis calculées sur les MGS, les fonctions KEGG potentielles ou exprimées. _____ 133
Figure F-5 : Distribution des abondances relatives les plus fortes en métagénomique et en métatranscriptomique. _____ 135
Figure F-6: Distribution des LFC entre métagénomes et métatranscriptomes. _____ 137
Figure F-7 : Analyse triadique partielle sans a priori sur la structuration des échantillons en fonction de la lignée et du régime._____ 138
Figure F-8: Distribution des corrélations entre métagénomes et métatranscriptomes. _____ 140
Figure F-9 : Catégories fonctionnelles structurant les métatranscriptomes. _____ 143
Figure F-10 : Extrait du métabolisme du méthane et fonctions différentiellement exprimées entre régimes._____ 146
Figure F-11 : Transporteurs ABC dont les fonctions sont différentiellement exprimées entre lignées au sein du régime CTR._____ 149
Figure F-12 : Fonctions du métabolisme du propionate différentiellement exprimées entre lignée. 150
Figure G-1 : Contributions aux développements bioinformatiques des outils utilisés pendant la thèse. _____ 155

Figure G-2: Analyses comparatives multi-omiques. _____ 157
Figure G-3 : Processus d'analyses des métatranscriptomes. _____ 159
Figure H-1 : Modèle représentant les voies directes et indirectes de l'influence de la génétique sur le microbiote et un phénotype. _____ 164

TABLE DES TABLEAUX

Tableau A-1 : Liste des catalogues de microbiotes intestinaux de poule. _____ 41

Tableau C-1 : Résumé de la composition en céréales des deux régimes alimentaires. _____ 52

Tableau E-1 : Mesures quantitatives et qualitatives des contigs._____ 100

Tableau E-2 : Mesures de longueurs d'assemblage par groupe lignée x régime._____ 101

Tableau E-3 : Résumé quantitatif de l'annotation structurale et fonctionnelle. _____ 102

Tableau E-4 : Mesure quantitative et qualitative des MGS. _____ 104

Tableau E-5 : Comparaison du catalogue MetaChick et du catalogue *de novo* enrichi._____ 106

Tableau E-6 : Comparaison des méthodes de quantification taxonomique appliquées au catalogue MetaChick._____ 108

Tableau E-7 : Nombre de gènes et de MGS du catalogue enrichi présents, quantifiés, filtrés. ____ 111

Tableau E-8 : Organismes hôtes des génomes appartenant aux genres et espèces détectés uniquement par co-assemblage. _____ 112

Tableau E-9 : Abondance relative et effet de la lignée et du régime à l'échelle des phyla._____ 115

Tableau E-10 : Nombre et pourcentage de gènes (quantifiés) annotés avec KEGG et/ou CAZy.___ 119

Tableau F-1 : Comparaison des effets de la lignée et du régime sur les indices de richesse et de β-diversité estimés sur les taxonomies ou les fonctions. _____ 132

Tableau F-2 : Nombre de fonctions KEGG différentiellement abondantes et/ou exprimées entre lignées ou entre régimes. _____ 134

Tableau F-3 : Coefficient RV entre métagénomes et métatranscriptomes._____ 139

Tableau F-4 : RDA : variance expliquée par les modèles contraints et significativité des variables explicatives, lignée et/ou régime._____ 141

Tableau F-5 : Nombre de fonctions DA et/ou DE parmi le top 5% des fonctions qui contribuent le plus à la structuration des métatranscriptomes en fonction de la lignée et du régime. _____ 142

A. INTRODUCTION

A.1. Contexte historique et agro-socio-économique de la poule pondeuse

A.1.1. Domestication et répartition mondiale de la poule domestique

Le genre *Gallus*, auquel appartient la poule domestique, est divisé en quatre espèces toutes originaires de l'Asie du Sud et du Sud-Est (**Figure A-1**). *Gallus gallus*, la poule dorée (« *Red Junglefowl* »), occupe une large zone s'étendant de l'Indonésie au centre de la Chine , et du nord-est de l'Inde aux îles des Philippines. *Gallus varius*, la poule de Java (« *Green Junglefowl* »), est sympatrique de *Gallus gallus* mais se limite à une région plus restreinte centrée sur l'île de Java. *Gallus sonnerati*, la poule de Sonnerat (« *Grey Junglefowl* »), est principalement présente en Inde; tandis que *Gallus lafayetti*, la poule de Lafayette (« *Ceylan Junglefowl* »), est endémique du Sri Lanka (Lawal and Hanotte 2021).

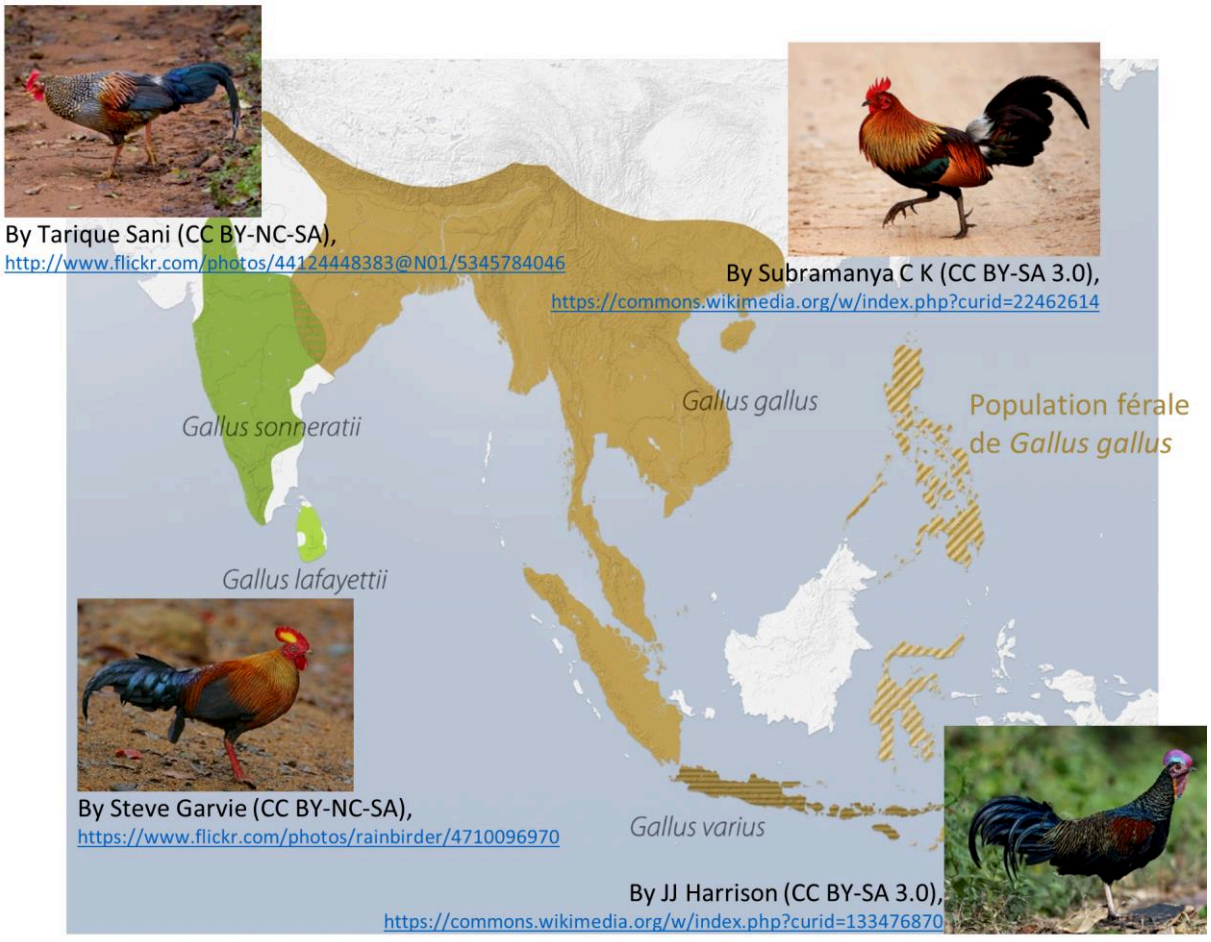


Figure A-1: Répartition des populations sauvages du genre Gallus.
Cartographie adaptée de Donkey shot 2016.

L'histoire de la domestication de la poule remonte au Néolithique, il y a environ 8 000 ans. La poule domestique, désignée scientifiquement sous le nom de *Gallus gallus domesticus*, descend principalement de la poule dorée, et dans une moindre mesure de la poule de Sonnerat (Tixier-Boichard, Bed'hom, and Rognon 2011; Mariadassou et al. 2021). La localisation exacte du début de cette domestication reste incertaine et il se pourrait qu'elle ait eu lieu à différents endroits de la zone de répartition de la poule dorée, même si de plus en plus d'études suggèrent une zone entre la Thaïlande, le Myanmar et la Chine. La domestication a conduit à la diffusion des poules domestiques à travers le monde en suivant les migrations humaines et les routes commerciales vers l'ouest et l'Europe avant d'atteindre le continent américain. Dans un premier temps, cette domestication n'a pas eu un but alimentaire, mais plutôt culturel. Désormais répartie globalement là où les humains vivent (M. Gilbert et al. 2018; 2022) (**Figure A-2**), l'espèce *Gallus gallus* s'illustre par sa capacité d'adaptation à des environnements très variés. Des régions sèches ou humides, avec des températures chaudes ou froides, que ce soit en altitude ou dans les plaines, on retrouve partout dans le monde des élevages de poules.

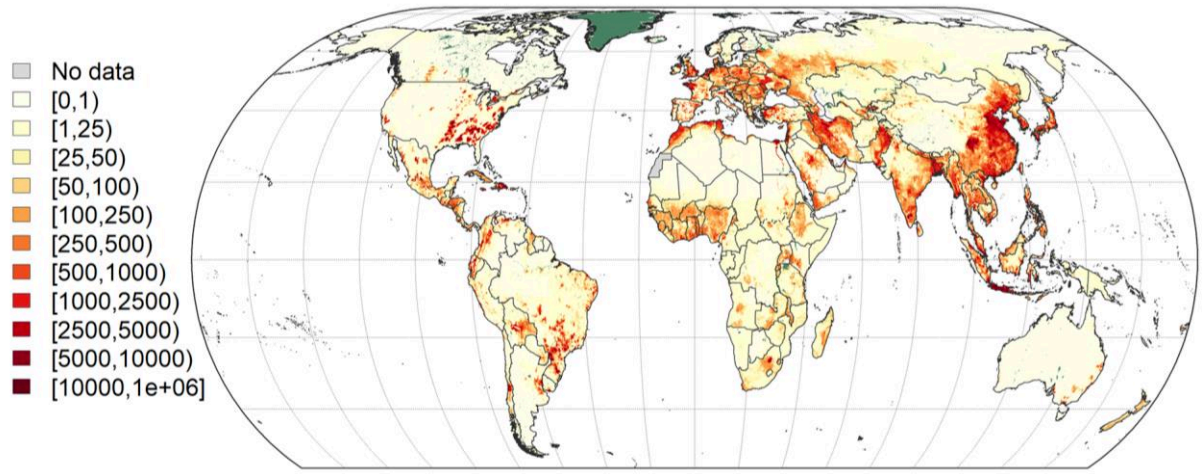


Figure A-2: Répartition du nombre poules d'élevage en 2015 à travers le monde.
Les zones en gris foncé sont ignorées par l'étude et les zones en vert foncé correspondent aux zones protégées par l'Union Internationale pour la Conservation de la Nature (IUCN). Figure extraite de M. Gilbert et al. 2022.

A.1.2. Pratiques d'élevage et importance des œufs en nutrition humaine

Les poules sont élevées dans trois grands types de systèmes caractérisés par leur taille, l'investissement nécessaire, et *in fine* leur production. Le système extensif et de petite taille, dans la continuité des débuts de la domestication, utilise essentiellement des animaux de souches locales, qui sont élevés en liberté et se nourrissent en autonomie, avec les déchets alimentaires ou à partir de cultures locales. Ces élevages sont appelés ruraux, familiaux ou de village. À l'opposé, les élevages

intensifs de grande taille nécessitent de forts investissements et permettent de contrôler l'environnement d'élevage. Ils utilisent des souches et une alimentation commerciale, améliorant ainsi la production. Entre ces deux extrêmes se trouvent les élevages semi-intensifs et semi-commerciaux, de tailles moyennes, qui mixent les pratiques d'élevage des deux systèmes précédents. Bien que la majorité de la production de poules et d'œufs provienne d'élevages commerciaux intensifs des pays développés, les élevages ruraux sont encore d'une grande importance dans les régions du monde en voie de développement, à faibles revenus et où la sécurité alimentaire peut faire défaut. Ainsi, au début des années 2000, Pym et al. (Pym, Bleich, and Hoffmann 2006) ont montré qu'en moyenne 80% des poules des pays africains et asiatiques à faible développement économique étaient des poules indigènes majoritairement élevées dans des systèmes traditionnels ruraux ou péri-urbains de petites échelles. Bien que ces animaux présentent des performances faibles, ils contribuaient de 20 à 30% des apports protéiques dans l'alimentation humaine de ces régions du monde (Wong et al. 2017). Depuis les pratiques d'élevage ont évoluées, mais ces élevages ruraux représentent toujours une ressource financière supplémentaire permettant de limiter la pauvreté (Birhanu et al. 2023), notamment pour les femmes qui ont généralement en charge ces élevages et qui rencontrent souvent, dans ces régions, de plus grandes difficultés à accéder aux différentes ressources (terres, travail, éducation) (Wong et al. 2017).

Que ce soit *via* des élevages à grande échelle commerciale, ou de village, les poules jouent un rôle majeur dans l'alimentation humaine. Les besoins nutritionnels humains se divisent en différents composants plus ou moins essentiels (c'est-à-dire synthétisable ou non par l'homme lui-même), en plus ou moins grande quantité (macro- ou micro-nutriments). Les nutriments essentiels incluent des acides aminés, des acides gras, des vitamines, des minéraux ou encore la choline. En comparaison aux aliments d'origine végétale, les aliments d'origine animale contiennent une majorité de ces nutriments et présentent l'avantage, d'être plus digestes et biodisponibles pour l'homme, *i.e.* non résistants et mieux absorbables (Chungchunlam and Moughan 2023; Day, Cakebread, and Loveday 2022; Gaudichon and Calvez 2021). En particulier, la digestibilité des protéines animales est estimée à 93% contre 80 % des protéines végétales (Gaudichon and Calvez 2021). De plus, la biodisponibilité des vitamines d'origine animale est souvent supérieure ou équivalente à celle des vitamines d'origine végétale (Chungchunlam and Moughan 2023). *A contrario*, les plantes contiennent moins d'acide gras saturés qui ont des effets négatifs sur la santé, et sont globalement plus riches en fibres, en vitamine C, en antioxydants et en certains minéraux que les aliments d'origine animale (Slavin and Lloyd 2012; Kromhout et al. 2016).

Parmi les aliments d'origine animale, les œufs ont un intérêt particulier en nutrition humaine. Un œuf est, en fait, un ovule (la plus grosse cellule du règne animal) qui, lorsqu'il est fécondé, doit pourvoir

aux besoins nutritionnels d'un embryon en développement sans apport extérieur, et assurer sa défense pendant quatre semaines. Il est composé, outre la coquille qui isole l'embryon de l'environnement extérieur, d'un compartiment jaune nourricier et d'un compartiment blanc défensif grâce à ses protéines antimicrobiennes (Nys and Guyot 2011). La composition combinée du jaune et du blanc de l'œuf répond à la majorité des besoins nutritionnels humains. En effet, on y trouve des protéines considérées de haute qualité, des minéraux, l'ensemble des vitamines essentielles à l'exception de la vitamine C, et des lipides (Iannotti et al. 2014). Le jaune est le compartiment qui contribue le plus à ces apports nutritifs, le blanc étant principalement composés de protéines. En particulier, la vitamine A est en concentration élevée dans le jaune d'œuf en comparaison aux autres aliments d'origine animale (FAO 2023a). La quantité des vitamines B2 et B12 contenue dans un seul œuf (~60g) couvre plus de la moitié et jusqu'à la totalité des apports journaliers recommandés pour l'être humain. L'œuf contribue également à l'apport de lipides essentiels, tout en limitant l'apport d'acides gras saturés contrairement aux autres sources d'aliments carnés (FAO 2023a; Nys and Guyot 2011; Réhault-Godbert, Guyot, and Nys 2019). Les œufs contiennent également des composants antinutritionnels qui peuvent interférer avec la digestion et l'absorption des nutriments, mais ils sont partiellement dégradés lors de la cuisson. Parmi ces composants, on trouve des protéases qui réduisent l'activité enzymatique, ou des protéines qui se lient aux vitamines ou aux minéraux et réduisent leur biodisponibilité. Pendant longtemps, les recommandations nutritionnelles étaient de limiter la consommation des œufs en raison de leur teneur élevée en cholestérol et donc leur potentielle association avec les maladies cardiovasculaires. Désormais, il est admis que celui des œufs n'influence pas le cholestérol sanguin, et l'œuf est désormais considéré comme un aliment aux nombreux bénéfices nutritionnels (Réhault-Godbert, Guyot, and Nys 2019).

Cette richesse nutritionnelle, associée à un coût abordable et à l'absence de contraintes cultuelles ou culturelles autour de sa consommation, fait de l'œuf une ressource alimentaire populaire et impliquant ainsi une production croissante.

A.1.3. Importance et évolution de la production d'œufs vis-à-vis des autres productions animales

La poule n'est pas la seule espèce d'oiseaux élevés pour la viande ou les œufs (FAO 2023b). Les cailles, canards, oies, dindes … sont également produits significativement par les industries avicoles (pour leur chair et/ou leurs œufs), mais les poules dominent le secteur, représentant en 2022 plus de 93% de ces oiseaux d'élevage. En comparaison avec d'autres types de productions animales (viande et œufs), les œufs de poule constituent la troisième source d'aliments d'origine animale produite avec près de 87

millions de tonnes d'œufs en 2022 (**Figure A-3**). Ils se situent derrière la viande de poulet (123,6 millions de tonnes) et la viande de porc (122,6 millions de tonnes), mais devant la viande de bœuf (69 millions de tonnes). La production mondiale d'œufs est en constante augmentation. Elle a notamment été multipliée par 2,5 depuis 1988, année à partir de laquelle l'Asie est devenue incontestablement le leader mondial de la production d'œufs, devant l'Europe. En effet, entre 1988 et 2022 l'Asie a quadruplé sa production.

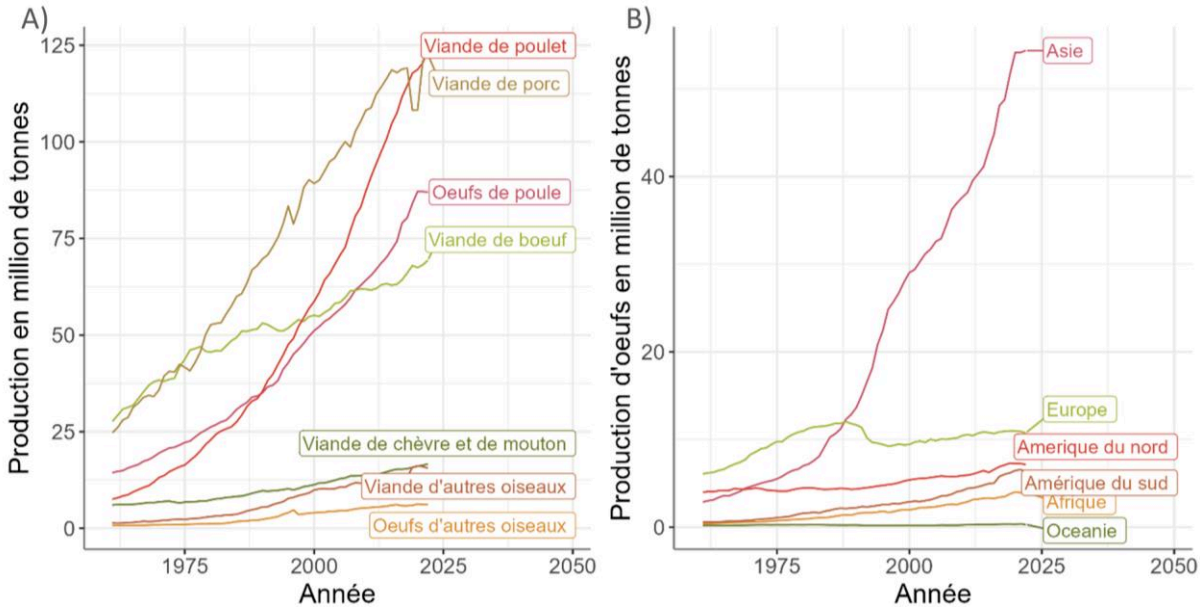


Figure A-3: Représentation de la production mondiale de viande et d'œufs.
A) en million de tonnes produits par type de produit, B) répartition de la production d'œufs de poules selon les régions du monde (FAO 2023b).

L'augmentation observée de la production avicole répond à une tendance générale de demande croissante pour ces produits. Cette demande est stimulée par la croissance démographique mondiale ainsi que par l'élévation générale du niveau de vie, qui influence directement la quantité de nourriture consommée (van Dijk et al. 2021). Compte tenu de ces facteurs, ainsi que de la baisse de la population souffrant de malnutrition, une augmentation de 30 à 62% de la demande en nourriture est attendue d'ici 2050. Cette prévision se manifeste clairement dans les statistiques concernant les produits d'origine animale avec, depuis 2010, une augmentation de 5% de la quantité de viande consommée par habitant dans le monde, et de 18% en ce qui concerne les œufs de poule (H. Ritchie, Rosado, and Roser 2024).

Ainsi l'élevage de poules et notamment des poules pondeuses, est porté par des intérêts économiques importants et croissants. Il joue également un rôle important dans la sécurité alimentaire mondiale en

améliorant les apports nutritifs nécessaires aux populations défavorisées et en leur fournissant un revenu complémentaire.

A.2. Facteurs d'évolution de la production d'œufs

Pour atteindre les niveaux actuels de production, d'importants développements ont été réalisés que ce soient à l'échelle des systèmes de production et des pratiques d'élevage, qu'au niveau de la composition des régimes alimentaires des animaux ou encore au niveau de la sélection de lignées animales plus performantes. Même si on considère que la plus grande partie des améliorations des performances sont dues au progrès génétique (Havenstein, Ferket, and Qureshi 2003), ces avancées sont intrinsèquement liées aux recherches menées en nutrition et en gestion des systèmes d'élevage. En effet, les développements en nutrition animale doivent s'adapter aux nouveaux besoins nutritionnels des animaux qui évoluent en fonction de l'âge, des lignées et des conditions d'élevage (Leclercq, Henry, and Lebas 1996). Inversement, la sélection génétique a permis la production de lignées animales capables de s'adapter à différents régimes alimentaires (Ito 2023).

A.2.1. La sélection génétique : focus sur l'efficience alimentaire

« L'efficience alimentaire consiste à évaluer la quantité d'aliments nécessaire pour obtenir une unité de produit animal (viande, œuf ou lait)» (Cantalapiedra-Hijar et al. 2020). Elle représente un caractère phénotypique d'intérêt majeur dans les schémas de sélection et ce chez toutes les espèces d'animaux d'élevage. La sélection intensive des animaux de rente a d'abord été motivée par des intérêts économiques en cherchant à augmenter la production en quantité et en qualité, ainsi qu'en cherchant à diminuer les coûts. Chez la poule, l'alimentation représente jusqu'à 70% des coûts de production (Neeteson-van Nieuwenhoven, Knap, and Avendaño 2013). Sélectionner des animaux pour qu'ils atteignent un fort niveau de performance (quantité et qualité de viande ou d'œufs, rapidité de croissance, …) en limitant les intrants, *i.e.* principalement la nourriture, résulte nécessairement d'une augmentation de revenus pour l'éleveur. Désormais, les enjeux environnementaux font partis des éléments pris en compte dans les schémas de sélection. L'efficience alimentaire permet de répondre à ces différents objectifs en se focalisant sur l'optimisation de l'utilisation des ressources.

Conceptuellement, la mesure de l'efficience alimentaire évalue la quantité d'aliment ingérée (« *Feed* Intake », FI) vis-à-vis de la quantité de produit (viande, lait ou œufs) obtenue. Une première mesure de l'efficience, l'indice de consommation (« *Feed Conversion Ratio* », FCR), correspond directement à

cette mesure en calculant le rapport entre ces deux quantités (**Figure A-4-A**). Chez la poule pondeuse, la quantité de produit est représentée par la masse d'œufs produits :

$$FCR = \frac{Quantité\ ingérée\ d'aliments(kg)}{Masse\ d'oeufs\ produits\ (kg)}$$

Plus ce rapport est petit, plus la nourriture utilisée pour nourrir les animaux est valorisée par la production d'œufs. Par sa simplicité de calcul, c'est l'indice qui est majoritairement utilisé par les entreprises de sélection. Cependant, il ne différencie pas les besoins nutritionnels physiologiques de l'animal de ceux nécessaires à la production.

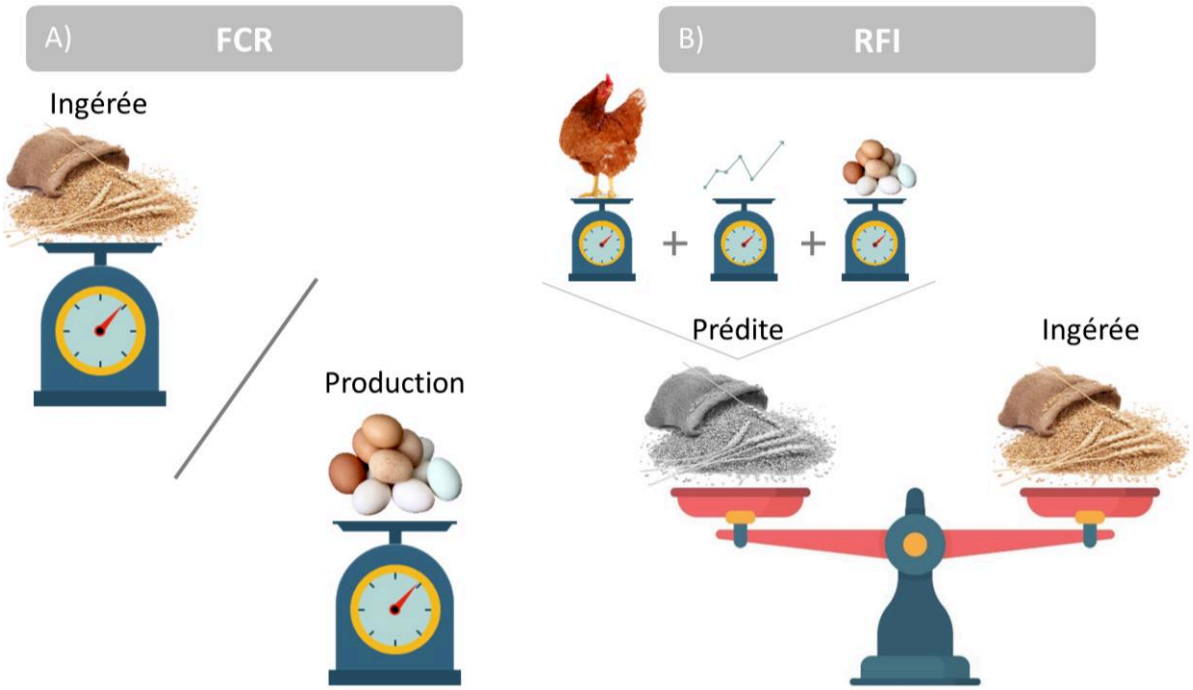


Figure A-4 : Représentation schématique des mesures d'efficience alimentaire.
A) L'indice de consommation, FCR, compare la quantité ingérée d'aliments à la quantité produite d'œufs; B) L'indice de consommation résiduelle, RFI, compare la quantité ingérée d'aliments à une quantité prédite qui tient compte du poids de la poule, du gain de poids et de la quantité d'œufs produite.

A contrario, l'indice de consommation résiduelle (« *Residual Feed Intake* », RFI) reflète la capacité innée de l'animal à utiliser efficacement les aliments, indépendamment du niveau de production. La RFI a été introduite par T.C. Byerly en 1941 chez la poule, comme la fraction de l'alimentation (quantité résiduelle) qui n'est ni utilisée pour les besoins d'entretien de l'animal, ni nécessaire à la production (Byerly 1941). Elle a ensuite été appliquée chez la vache en 1963 (Koch et al. 1963). Dans les années

70 A. Bordas et P. Mérat ont étudié comment la RFI pouvait être utilisée comme un critère de sélection pour améliorer l'efficience alimentaire des poules pondeuses et ont mis en place une sélection divergente pour ce caractère à partir de 1976 (Bordas and Mérat 1974; Bordas, Tixier-Boichard, and Merat 1992).

Le principe de cette mesure de l'efficience alimentaire est de comparer, sur une période donnée, la FI réelle à une quantité ingérée estimée (« *Predicted Feed Intake* », PFI) (**Figure A-4-B**). La PFI est calculée *via* un modèle de régression linéaire multiple qui prend en compte des mesures phénotypiques d'entretien (l'énergie nécessaire pour maintenir les fonctions physiologiques de base) et de production. Chez la poule pondeuse, le poids de l'animal et le gain de poids représentent les besoins d'entretien, et la masse d'œufs représente les besoins nécessaires à la production :

$$PFI = a * Poids\ de\ l'animal^{0.5} + b * Gain\ de\ poids + c * Masse\ d'oeufs\ produits + d$$

$$RFI = FI - PFI$$

Ainsi, si la RFI est négative, l'animal aura consommé moins que prévu pour son maintien et son niveau de production ; il sera donc efficient. Au contraire, si la RFI est positive, alors l'animal aura consommé plus qu'estimé et sera donc non efficient.

Cet indice de l'efficience a ensuite largement été utilisé en recherche chez différentes espèces d'élevage, et, bien que les modèles soient différents selon les espèces (notamment sur le paramètre relatif à la production), de multiples études ont montré une héritabilité modérée du caractère illustrant sa possible utilisation en sélection génétique. L'étude de la RFI dans différentes lignées sélectionnées ou non, a montré des corrélations phénotypiques positives ou négatives avec de nombreux autres caractères de production et physiologiques. Elle est en particulier, systématiquement corrélée à la prise alimentaire et au FCR (Tixier-Boichard et al. 2002). Chez la poule, elle a été observée comme fortement négativement corrélée avec la teneur en gras (les poules efficientes stockant plus de gras (El-Kazzi et al. 1995a)), et au contraire chez le porc, elle n'était pas corrélée avec l'épaisseur du gras dorsal (H. Gilbert et al. 2006). Cependant, les conditions de sélection ne sont pas les mêmes, car chez la poule, la sélection est faite à l'âge adulte, alors que chez le porc, elle est effectuée sur des animaux en croissance. L'interprétation de l'efficience alimentaire doit donc être contextualisée non seulement à l'espèce prise en compte mais également à la race (ou souche), au sexe ou à l'âge des animaux évalués. Chacun de ces paramètres influence fortement le modèle et démontre des mécanismes variables de l'efficience alimentaire selon les contextes. Par exemple, l'héritabilité chez des animaux plus jeunes (en croissance) tend à être plus faible que celle des animaux adultes (comme les poules

pondeuses, donc matures) (Tixier-Boichard et al. 2002), et la sélection de la RFI à l'âge adulte ne présage pas nécessairement d'une différence d'efficience chez le jeune. Par ailleurs, chez la poule pondeuse, la corrélation génétique entre les mesures d'efficience des mâles et des femelles sélectionnés sur une faible ou une forte efficience n'est pas différente de 0, révélant des mécanismes génétiques différents entre les deux sexes (Tixier-Boichard et al. 1995).

Enfin, l'efficience alimentaire est influencée par différents facteurs environnementaux, en particulier la température et la composition du régime alimentaire. Il a été observé que les animaux exposés à une températures ambiante élevée ont tendance à diminuer la RFI. Ceci est dû au fait que l'exposition à la chaleur induit une réduction significative de la prise alimentaire, ce qui conduit à une perte de poids corporel et de production qui peut être limitée au début mais qui peut devenir significative si le stress persiste (Labroue et al. 1999; Bordas and Minvielle 1997). En ce qui concerne le régime alimentaire, on peut noter que la teneur en énergie influence également la RFI. Une réduction de l'énergie a tendance à provoquer une augmentation de la prise alimentaire et donc une augmentation de la RFI (Bordas et al. 1995)).

A.2.2. Le régime alimentaire : un facteur clé de la production

Dans un contexte de production animale, la composition de l'alimentation, au-delà de subvenir aux besoins de l'animal pour sa croissance et sa santé, est un facteur clé d'optimisation des caractères de production. Comme pour un être humain, les recommandations en nutrition se divisent en apport en énergie, en acides aminés, et dans une moindre mesure en vitamines et en minéraux. Chez les animaux d'élevage, ces recommandations varient selon la lignée, le sexe, l'âge ou le stade de production et les conditions d'élevage. Pour évaluer l'apport en particulier en protéine et en énergie, des progrès significatifs ont été réalisés dans la caractérisation des nutriments présents dans chaque composant alimentaire, ainsi que dans la compréhension de leur digestibilité et de leur biodisponibilité (Ravindran 2012), et ce pour différentes espèces animales.

Les poules sont presque exclusivement nourries à partir de céréales dont la composition varie selon les régions du monde. La céréale principale utilisée pour subvenir aux besoins énergétiques est le maïs (qui favorise également la coloration du jaune), ou le blé (plus localement comme en France) ou certaines variétés de sorgho (notamment dans les pays chauds). Le soja sert de source principale de protéines, qui localement peut être remplacé par le canola, le pois ou le tournesol (Ravindran 2013). En plus des caractéristiques intrinsèques de chacune des céréales, les différents processus de transformation de cette matière première ont un impact sur leurs valeurs nutritives et sur leur

digestibilité. Par exemple, la taille des ingrédients ne doit être ni trop fine ni trop grossière pour qu'ils soient correctement assimilés par les poules (Ito 2023) ; et les graines doivent être décortiquées pour réduire la proportion de fibres indigestes, et éventuellement chauffer pour détruire les composants antinutritionnels comme des protéases. *In fine*, il s'agit d'associer ces différentes formes de céréales pour atteindre de façon équilibrée et digeste les besoins de l'animal. En effet, au-delà de la digestibilité de chaque composant individuellement, le régime alimentaire doit être digeste dans son ensemble. Son contenu en protéine, en lipide, en amidon et en fibres (selon leur origine et leur quantité) influence la digestibilité d'autres composants et donc potentiellement les performances des animaux (Selle and Liu 2019; Desbruslais et al. 2021). Par exemple, les fibres sont globalement considérées comme antinutritionnelles, mais il est désormais communément admis d'avoir 3% de fibres non soluble dans le régime alimentaire pour permettre de fluidifier le contenu intestinal et améliorer l'absorption des nutriments par la poule (Tejeda and Kim 2021). Finalement, la formulation du régime peut être complétée par des minéraux, des acides aminés ou encore des enzymes. Le calcium et le phosphore sont particulièrement important chez la poule pondeuse car ils interviennent dans la formation des os de la poule et de la coquille de l'œufs. En cas de déficit de ces minéraux dans l'alimentation, la qualité des œufs diminue (coquille plus fine, cassante ou molle), et la poule puise dans ces propres réserves, notamment ses os, pour combler les besoins de constitution de la coquille, ce qui peut engendrer des fragilités osseuses importantes (Lehr 2021; Sinclair-Black, Garcia, and Ellestad 2023). La méthionine et la lysine sont des acides aminés particulièrement important car ils sont en déficit dans le régime classique à base de maïs et soja et qu'ils influencent la production des œufs (en taille ou en composition) (Macelline et al. 2021). Enfin, les enzymes permettent en particulier d'améliorer la dégradation des ingrédients notamment ceux considérés comme moins nutritifs car plus riches en fibres. En effet, les fibres solubles augmentent la viscosité du contenu intestinal et réduit la digestibilité et l'absorption des nutriments (Alagawany, Elnesr, and Farag 2018).

Ces connaissances en nutrition permettent de formuler des régimes optimaux, qui répondent à la fois aux besoins de l'animal et aux objectifs de production, ou des régimes alternatifs qui prennent également en compte la disponibilité et le coût des ingrédients.

A.2.3. Evolutions des performances et mise à jour des objectifs de recherche et développement

Chez la poule, l'amélioration génétique des lignées et l'optimisation de la composition de l'alimentation et des pratiques d'élevage ont permis d'améliorer les caractères de production tout en réduisant la prise alimentaire. La sélection a permis d'atteindre (i) un début de pic de ponte plus

précoce, (ii) un allongement de la durée de ponte à qualité constante, et (iii) un poids de l'œuf constant et une préservation de la qualité de la coquille sur la période de ponte (Bain, Nys, and Dunn 2016). Ainsi, Anderson et al. ont noté en 2009, que les poules pondeuses de deux lignées commerciales arrivaient à maturité sexuelle autour de la 17^{e} semaine et atteignaient 50% du taux de production autour de la 20^{e} semaine, ce qui était 1 mois plus tôt comparé à 1958 (Anderson et al. 2013; Hy-line 2024). Désormais les poules pondeuses peuvent être maintenues en production jusqu'à la 100^{e} semaine, avec un taux de production maximale proche de 100%, et ce maintenu sur une longue période (au-delà de la 65^{e} semaine, et >80% au-delà de la 80^{e} semaine) (van Sambeek 2010; lohmann 2016; Hy-line 2024). Sur le cycle annuel d'une poule, la production totale dépasse 300 œufs et peut atteindre 350 œufs, soit une augmentation de plus de 30% par rapport à 1958 (Preisinger 2018; Shcherbatov and Shkuro 2021; Korver 2023). Face à un taux de ponte qui approche du maximum possible, les efforts se concentrent désormais sur l'allongement de la durée de ponte des poules. Du point de vue de l'efficience alimentaire, Anderson et al. rapportent une baisse de 45 à 50% des valeurs de FCR entre 1958 et 2009 (Anderson et al. 2013) sur deux lignées commerciales de poules pondeuses ; de même, Preisinger indique qu'entre 1995 et 2015, la quantité de nourriture pour produire 1kg d'œuf a diminué de 450g (Preisinger 2018). *In fine*, ces études indiquent qu'il faut environ 2kg d'aliments pour produire 1kg d'œuf (contre 1,3 à 3,2kg pour la viande de poulet, 3 à 3,9kg pour la viande de porc, ou 8 à 12kg pour la viande de bœuf (Pradeep 2020; Elem 2024)).

Répondant initialement à des enjeux économiques, les développements sont de plus en plus motivés par des questions d'adaptation à des environnements d'élevage changeant et à des contraintes sociétales nouvelles (Akinyemi and Adewole 2021). Parmi ces nouvelles motivations, il s'agit de mieux prendre en compte le bien-être animal dans les élevages (avec notamment l'arrêt des élevages en cage en Europe) ou bien de réduire l'empreinte environnementale des élevages. Pour cela, la recherche en nutrition animale diversifie les sources d'ingrédients, comme les co-produits des transformations céréalières par ailleurs utilisées pour l'alimentation humaine, des algues ou encore des insectes. Des études sont également menées sur l'introduction de nouveaux compléments alimentaires et de nouvelles formulations de régimes afin de limiter les apports nutritifs excessifs qui sont rejetés ensuite devenant alors sources de pollution (Macelline et al. 2021; Ito 2023). L'efficience alimentaire permet également de répondre à ces enjeux de réduction d'impact sur l'environnement. Un animal plus efficient qui consomme la juste quantité de nourriture pour ses besoins physiologiques et pour les objectifs de production, réduit la quantité d'intrant ainsi que les déchets.

L'accès à un régime alimentaire optimal est difficile dans beaucoup de régions du monde du fait de son coût et ses fluctuations, notamment à cause de contextes géopolitiques instables (Wong et al. 2017). Par ailleurs, le secteur de l'alimentation animale se retrouve en compétition pour l'accès à

certaines céréales avec d'autres filières, notamment l'alimentation humaine. De plus, la réduction de l'empreinte environnementale des élevages peut restreindre la composition des régimes alimentaires par exemple aux céréales cultivées à proximité (Leinonen and Kyriazakis 2016). Enfin, la production des lignées commerciales est de plus en plus mondialisée et les animaux de production sont confrontés à des environnements très variables. Ainsi, la compréhension des mécanismes permettant l'extraction complète des nutriments et leur valorisation de manière efficiente, reste un domaine actif de recherche pour répondre à ces contextes variés et changeants.

Pour répondre à cet enjeu, de plus en plus d'études s'intéressent au rôle que pourrait avoir le microbiote intestinal dans l'efficience alimentaire des animaux d'élevage.

A.3. Le microbiote intestinal : généralités et focus sur la poule pondeuse

A.3.1. Microbiote : Définitions et interactions avec l'hôte

Le microbiote, dérivé du grec "mikros", et "biotos", qui signifient « petit » et « vie », définit l'ensemble des micro-organismes qui composent un écosystème. Ces micro-organismes incluent les procaryotes (bactéries et archées), mais également quelques eucaryotes comme des champignons, des levures, ou des protistes. Les recherches sur le microbiote se sont largement développées à partir des années 2000, en particulier pour caractériser les interactions possibles entre ces micro-organismes et le milieu dans lequel ils se trouvent (hôte vivant ou environnement naturel). Le terme microbiome, défini pour la première fois par Whipp et al. en 1988, comprend l'analyse du microbiote dans un milieu clairement défini, des fonctions qu'il réalise, ainsi que de l'ensemble des interactions entre les micro-organismes, et entre micro-organismes et leur environnement (Berg et al. 2020). La notion d'interaction est un concept fort qui permet notamment de classer les micro-organismes d'un microbiote selon qu'ils interagissent ou non avec l'hôte ou avec d'autres micro-organismes, et que ces interactions soient bénéfiques ou préjudiciables (Lederberg and McCray 2001). Ainsi, si les interactions sont positives ou neutres, on parle de micro-organismes mutualistes ou commensaux, selon que les deux organismes tirent profit ou non de cette interaction. *A contrario*, si les interactions sont négatives, on parle de micro-organismes parasites voire pathogènes selon l'impact sur l'hôte. Enfin, l'hôte et ses microbiotes peuvent être considérés comme une entité unique que l'on appelle holobionte, et de nombreuses études étudient les relations de co-dépendance et de co-évolution entre les éléments qui le composent (Bordenstein and Theis 2015; Simon et al. 2019; Groussin, Mazel, and Alm 2020).

Au cours des dernières années, les recherches sur les microbiotes ont été menées dans de multiples domaines, dans divers environnements naturels, chez l'homme et les animaux. Ces études ont permis

de révéler des associations entre communautés microbiennes et des phénotypes variés. A l'échelle d'écosystèmes environnementaux, on peut citer le projet Tara Ocean Microbiome (Fondation Tara Océan 2020) qui vise à décrire les microbiotes des océans en lien avec le dérèglement climatique et la pollution, ou bien les analyses sur les interactions entre micro-organismes et système racinaire qui influencent l'absorption de minéraux par les plantes (Jacoby et al. 2017). Chez l'homme, les microbiotes de différents organes ont été caractérisés notamment pour leurs potentielles associations avec la santé. Par exemple, le Human Microbiome Project (Human Microbiome Project Consortium 2012) est un projet ambitieux qui continue de décrire le microbiote intestinal, oral, nasal, de la peau ou encore vaginal, et qui a permis de mettre en évidence des associations entre dysbiose[*] des microbiotes et troubles fonctionnels (syndrome du côlon irritable), maladies inflammatoires (maladie de Crohn), pathologies cutanées, *etc.* (Pimentel and Lembo 2020; Carmona-Cruz, Orozco-Covarrubias, and Sáez-de-Ocariz 2022; Núñez-Sánchez et al. 2022). Par ailleurs, le microbiote intestinal n'influence pas que son environnement proche, il a également été associé à d'autres maladies liées à des organes plus éloignés comme le diabète ou des maladies mentales comme l'autisme (Sadagopan et al. 2023; Grau-Del Valle et al. 2023). Ces interactions hôte-microbiote sont étudiées sous la forme d'axes entre organes : cerveau-intestin-microbiote, ou encore foie-intestin-microbiote, et illustrent les échanges à double sens entre l'hôte et ses microbiotes (Clark and Mach 2016; Dinan and Cryan 2017; Albillos, de Gottardi, and Rescigno 2020). Ces axes d'interactions sont également étudiés chez les animaux d'élevage en particulier l'axe cerveau-intestin-microbiote (Kraimi, Dawkins, et al. 2019). Même si la causalité de ces interactions n'est pas toujours clairement établie, ces études ont mis en évidence des associations entre le microbiote intestinal et des phénotypes différenciés de stress, de mémorisation, ou de comportements sociaux ou alimentaires, chez le cheval (Destrez et al. 2015), différentes volailles (Birkl et al. 2018; Kraimi, Calandreau, et al. 2019), le porc (Val-Laillet et al. 2017) ou encore le bovin (DeVries and Chevaux 2014). L'impact du microbiote sur des phénotypes de santé des animaux d'élevage est également largement étudié notamment pour empêcher ou limiter la prolifération de bactéries pathogènes et en vue de réduire l'utilisation d'antibiotiques, notamment chez le porc et la poule (Fouhse, Zijlstra, and Willing 2016; Awad, Hess, and Hess 2018; Argüello et al. 2019; Khan et al. 2020; Peng et al. 2021; Horvathova et al. 2023). Enfin les études sur les microbiotes des animaux d'élevage illustrent les liens possibles avec les performances notamment de production, ou l'impact sur l'environnement. Par exemple, la production de méthane par les ruminants est notamment associée à l'activité de fermentation des carbohydrates par le microbiote ruminal (Tapio et al. 2017), des associations existent entre microbiote intestinal et des caractères de qualité du lait chez la vache (Xue et al. 2020), de prise de poids et d'efficience alimentaire chez le lapin (Velasco-Galilea et al. 2021),

[*] dysbiose : déséquilibre dans la composition des communautés de micro-organismes d'un écosystème.

de performance d'endurance chez le cheval de course (Plancade et al. 2019) ou d'efficience digestive contrastée chez le poulet de chair (Borey et al. 2020).

A.3.2. Les fonctions et productions du microbiote intestinal

Même si des microbiotes d'origines très variées sont de plus en plus étudiés, le microbiote intestinal est le mieux caractérisé. Chez la poule comme chez les autres animaux, le microbiote intestinal occupe des fonctions digestives et métaboliques, immunitaires et de défense, et neurologiques. En effet, il pourrait interagir avec le système nerveux central *via* des interactions avec les nombreux neurones qui composent le système nerveux entérique qui contrôle le système digestif (parfois appelé « le second cerveau ») (Kraimi, Dawkins, et al. 2019). Les fonctions digestives du microbiote sont essentielles car elles participent à la nutrition de l'hôte en dégradant les aliments qu'il n'est pas en mesure de digérer seul, et en favorisant l'absorption des nutriments. Elles permettent, entre autres, la production de nutriments parfois essentiels, comme des acides aminés et des vitamines (en particulier les vitamines K et B), et des acides gras à chaîne courte (« short-chain fatty acids », SCFA) (Shang et al. 2018). Ces SCFA sont particulièrement intéressants car ils participent à la bonne santé de l'hôte et influencent de multiples tissus. Ils sont rapidement absorbés et contribuent, chez la poule, jusqu'à 8% des besoins énergétiques de l'animal (10% chez l'homme et de 70 à 80% chez les ruminants (McNeil 1984; Józefiak, Rutkowski, and Martin 2004; Liu et al. 2022)). Les trois principaux SCFA produits par le microbiote – acétate, butyrate et propionate – dérivent de la fermentation des carbohydrates indigestes. Ils représentent plus de 90% des acides gras produits par le microbiote cæcal. Les autres acides gras sont des acides gras ramifiés qui dérivent de la dégradation des protéines (Hussein et al. 2023; Jha et al. 2019). L'acétate est généralement l'acide gras majoritaire car produit par la plupart des bactéries anaérobies (Louis and Flint 2017). Il est métabolisé dans de nombreux tissus comme source d'énergie et stimule la production d'acides gras dans le foie ou les tissus adipeux (Morrison and Preston 2016; Rivière et al. 2016). Le propionate et le butyrate sont produits dans des proportions variables qui dépendent fortement de la composition du régime alimentaire (Lei et al. 2012). Les bactéries qui les produisent sont plus spécifiques à l'un ou l'autre de ces SCFA, avec une dominance des Bacteroidetes pour la production de propionate et de Bacillota pour celle du butyrate (Kircher et al. 2022). Le propionate, transporté jusqu'au foie, influence le métabolisme du glucose et stimule plus précisément la néoglucogenèse. Le butyrate est la source principale d'énergie des cellules épithéliales. Il favorise leur renouvellement, stimule la production de protéines de jonction et la production de mucus (Morrison and Preston 2016). Il joue donc un rôle majeur dans l'intégrité de la barrière intestinale et dans la lutte contre les infections pathogènes. Par ailleurs, la production de SCFA provoque une

diminution du pH, ce qui contribue également à limiter la prolifération de bactéries pathogènes. L'influence du microbiote intestinal sur la santé s'illustre aussi par son rôle dans la maturation des cellules immunitaires et dans la modulation de leur réponse lors d'une infection (Khan et al. 2020; Kogut, Lee, and Santin 2020) et par la production de bactériocines qui limitent directement la croissance de certaines populations bactériennes. Enfin, il représente une barrière en soi, par relations compétitives pour la consommation de nutriments ou pour l'adhésion à la paroi intestinale (Khan et al. 2020; Cantu-Jungles and Hamaker 2023).

La présente étude portant sur l'efficience alimentaire et sur l'adaptation à une modification du régime, notre intérêt s'est porté ici sur le rôle du microbiote dans la nutrition de la poule en raison de son implication dans ces deux phénotypes.

A.3.3. L'appareil digestif de la poule

Chez la poule, le tube digestif est composé du jabot, du proventricule et du gésier, du petit intestin composé de trois segments (duodénum, jéjunum, iléon), des cæca, du colon et du cloaque (**Figure A-5)** (Stanley, Hughes, and Moore 2014; Oakley et al. 2014; Shang et al. 2018; Khan et al. 2020).

Le jabot est essentiellement une poche qui sert de réserve et influence la satiété. Quand il se vide, cela donne le signal à l'animal de manger à nouveau. C'est dans le proventricule et dans le gésier que la dégradation des aliments commence. Cette dégradation fait appel à des mécanismes enzymatiques et mécaniques, le gésier étant un muscle qui, grâce aux petites pierres ingérées par la poule, réduit les aliments en un substrat plus homogène. Ce digesta passe ensuite dans le petit intestin où la digestion se poursuit. Il est le segment principal d'absorption des nutriments. C'est en particulier dans le duodenum que les enzymes pancréatiques et la bile hépatique sont sécrétées, notamment pour la digestion des protéines et des lipides. L'absorption des acides aminés, des acides gras et des vitamines a ensuite majoritairement lieu dans les segments suivants, le jéjunum et l'iléon. A la jonction entre l'intestin grêle et le gros intestin se trouvent les cæca, deux grands et fins réservoirs où a lieu l'absorption de l'eau, le traitement de l'urée, et surtout la fermentation des aliments résistants, c'est-à-dire non dégradés par l'hôte dans les segments précédents. Le tube digestif se termine ensuite par le colon (qui contrairement à son autre nom, le gros intestin, est très petit chez les oiseaux) et le cloaque (là où se rejoignent les système digestif, urinaire et génital).

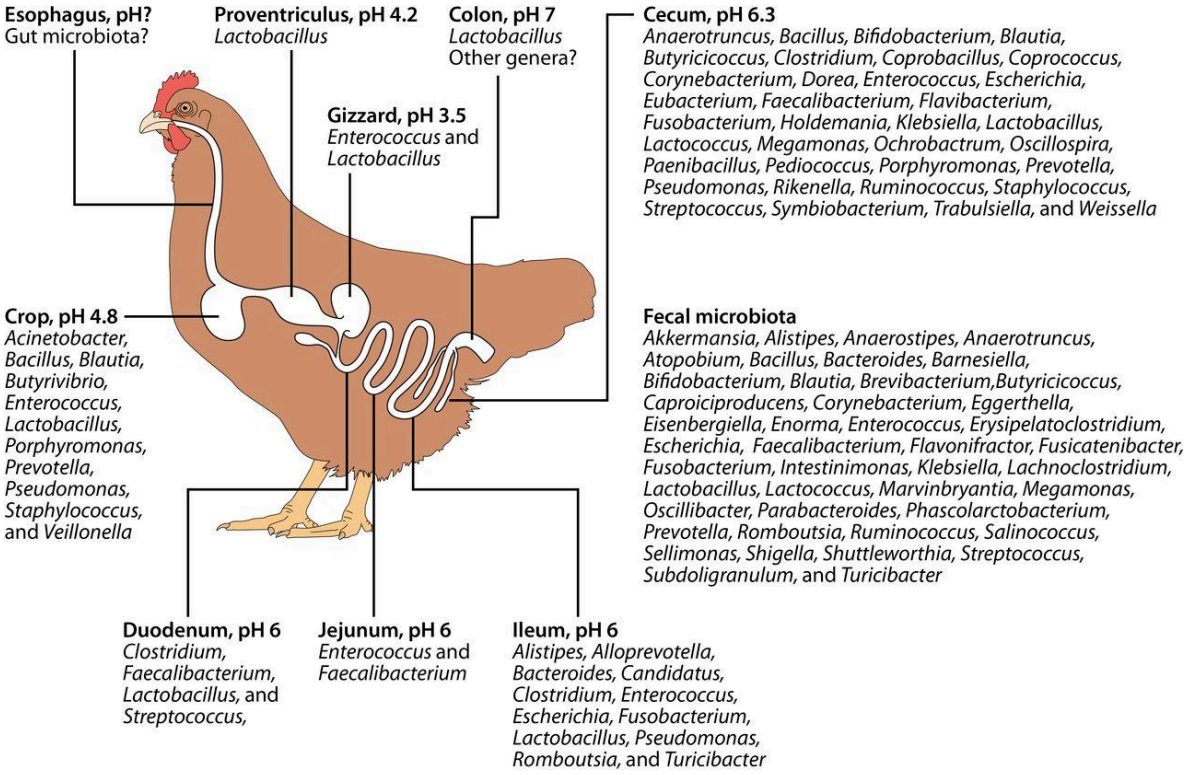


Figure A-5: Description du tractus digestif de la poule et principaux genres bactériens du microbiote intestinal.
Figure extraite de Khan et al. 2020.

A.3.4. Composition du microbiote intestinal le long du tractus digestif de la poule

Comme dans n'importe quel environnement des micro-organismes occupent les différents segments du tractus digestif de la poule, mais, du fait de propriétés physiques et chimiques variables, la composition du microbiote change selon les segments (**Figure A-5**). Le microbiote intestinal de poule pondeuse est presque exclusivement composé de bactéries (>98% (Glendinning et al. 2020; Plaza Oñate et al. 2023). La partie haute de l'intestin (du jabot au gésier) est considérée comme faiblement peuplée, notamment à cause du pH très acide (Oakley et al. 2014; Stanley, Hughes, and Moore 2014). Le nombre de cellules bactériennes augmente ensuite progressivement le long de l'intestin grêle jusqu'à atteindre un maximum au niveau des cæca (Shang et al. 2018).

La plupart des études sur le microbiote intestinal de la poule a été réalisée sur le contenu cæcal, bien que la majorité de l'absorption des nutriments ait lieu dans le petit intestin. Le rôle du microbiote au niveau du petit intestin est incertain, il pourrait jouer sur la disponibilité et l'absorption des nutriments, mais il est également considéré en compétition directe avec l'hôte pour la consommation d'énergie et de protéines (Stanley, Hughes, and Moore 2014; Shang et al. 2018). Par ailleurs, la vitesse de transit dans ce segment rend les prélèvements difficiles et les résultats montrent une variabilité importante selon l'échantillonnage (Kers et al. 2018; Shang et al. 2018). Au contraire, le cæcum est le segment où

le digesta demeure le plus longtemps, entre 12 et 20 heures contre environ 3,5 heures pour l'ensemble des autres compartiments (Pan and Yu 2014). Cette rétention favorise une plus grande stabilité des communautés microbiennes installées et une densité plus élevée de cellules microbiennes. Le microbiote cæcal présente également une plus grande richesse* et une plus grande diversité**.

Dans la partie haute de l'intestin et dans le petit intestin, bien que la richesse augmente au fur et à mesure, le microbiote est fortement dominé par le genre *Lactobacillus* (Videnska et al. 2013; Yan et al. 2017), alors que le cæcum contient en plus, de nombreuses autres espèces (**Figure A-6-A**), comme celles des genres *Bacteroides*, ou *Barnesiella* appartenant au phylum Bacteroidota, ou celles des genres *Ruminoccoccus*, ou *Faecalibacterium* appartenant au phylum Bacillota. Les phyla Actinomycetota et Pseudomonadota sont également plus représentés dans les cæca (Khan et al. 2020; Videnska et al. 2013; Yan et al. 2017; Rychlik 2020). Par ailleurs, c'est dans les cæca qu'a principalement lieu l'activité de fermentation microbienne des aliments non digérés par la poule, et là où le microbiote contribue le plus à la production de métabolites bénéfiques pour l'hôte (Stanley, Hughes, and Moore 2014). Comme décrit précédemment, les SCFA qui en découlent servent de sources d'énergie pour de multiples types cellulaires, influencent le métabolisme du glucose et des lipides, et participent au maintien de la santé de l'hôte en augmentant l'intégrité de la barrière épithéliale. Compte tenu de l'importance du rôle des cæca dans le développement d'un microbiote dense et riche, dans la fermentation et la production de métabolites tels que les SCFA, de nombreuses études font l'hypothèse que le microbiote cæcal pourrait exercer une influence sur l'efficience alimentaire des animaux (Gardiner, Metzler-Zebeli, and Lawlor 2020; Fang et al. 2020; He et al. 2023).

Les fèces sont généralement considérées comme un bon proxy de la composition du microbiote intestinal chez différentes espèces hôte, en particulier chez l'homme. Pour la plupart des Mammifères, les fèces présentent l'avantage d'être facilement prélevables, ne nécessitent pas le sacrifice des animaux et leurs compositions microbiennes restent globalement stables au cours du temps (Stanley, Hughes, and Moore 2014). Au contraire, chez la poule, le microbiote fécal est rarement analysé. En effet, son système digestif présente de grandes différences avec les autres animaux monogastriques. La vitesse de transit et la fréquence de vidange des différents compartiments est plus importante, la taille du gros intestin est particulièrement petite, et les urines sont excrétées sous forme solide avec les selles par le cloaque. Les analyses du microbiote fécal de poule ont ainsi montré une grande différence de composition avec celle des autres compartiments intestinaux et sa composition présente une variabilité importante dans le temps (c'est-à-dire entre prélèvement). Elle est majoritairement influencée par la vidange des différents compartiments intestinaux, les fèces étant soit de type cæcal

* Richesse : Nombre d'espèces différentes
** Diversité : Nombre d'espèces pondéré par leurs abondances

(vidange des cæca), soit de type fécal (vidange des autres segments intestinaux) (Sekelja et al. 2012; Pauwels et al. 2015; Stanley et al. 2015). Les fèces ne représentent donc pas un bon proxy du microbiote intestinal chez la poule. Notre étude portera donc également sur l'analyse des contenus cæcaux.

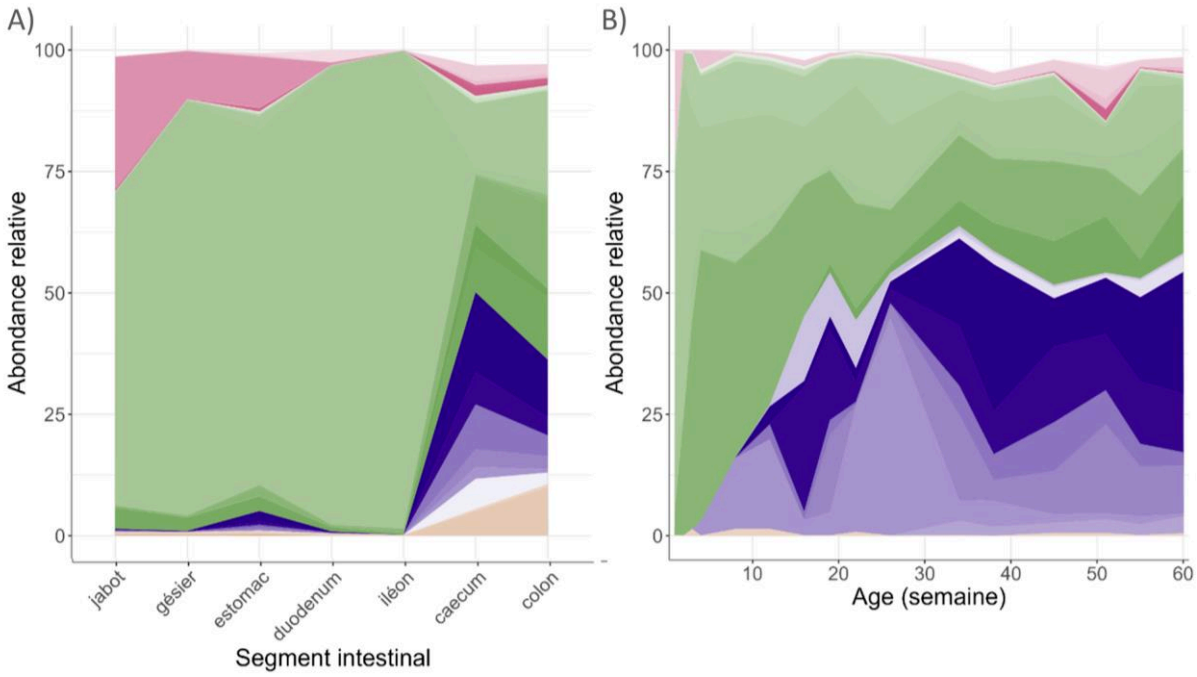


Figure A-6: Abondances relatives des familles des quatre phyla majoritaires du microbiote intestinal de poule pondeuse.
A) le long du tractus digestif à l'âge adulte (46 semaines) ; B) du cæcum en fonction de l'âge. Chacune des 106 familles est colorée selon son Phylum : en rose les Pseudomonadota, en vert les Bacillota, en violet les Bacteroidota, en orange les Actinomycetota. Figure adaptée de deux études : Videnska et al. 2013; 2014. A noter que les noms des segments intestinaux ont été reportés fidèlement à la publication de Videnska, un doute persiste quant à l'organe « estomac ». Il pourrait s'agir du proventricule mais dans ce cas il devrait être positionné avant le gésier.

A.3.5. Principaux facteurs de variation du microbiote intestinal de poule

Au-delà du choix du segment intestinal pris en compte, de multiples facteurs sont responsables de variations de la composition du microbiote (Kers et al. 2018; Khan et al. 2020). Certains sont liés aux caractéristiques de l'hôte (âge, sexe, lignée ou souche génétique), et d'autres à des facteurs environnementaux (système d'élevage, régime alimentaire, *etc.*). Comme précisé en première partie de cette introduction, ces facteurs peuvent également influencer la productivité des animaux. De façon logique, la composition du régime alimentaire a une forte influence sur la composition du microbiote cæcal (Kogut 2022). Celui-ci étant le lieu de fermentation des composés résistants au système digestif de la poule, utiliser une composition alimentaire plus ou moins digeste (en jouant notamment sur les

fibres) va nécessairement influencer la composition et la quantité de substrat atteignant les cæca, et permettre la croissance de micro-organismes variés. Par ailleurs, le type de fibres et leur origine (la céréale à l'origine de la fibre) vont également influencer la composition du microbiote. En effet, chaque micro-organisme ne dispose pas de l'ensemble des enzymes permettant la dégradation de toutes les fibres et celles-ci sont plus ou moins efficaces en fonction de la variante des fibres (Tiwari, Singh, and Jha 2019; Cantu-Jungles and Hamaker 2023). Ainsi, la quantité et la diversité des céréales inclues dans le régime vont influencer la quantité et le type de fibres atteignant les cæca et, *in fine*, moduler la composition de son microbiote.

L'âge couplé à l'environnement d'élevage est un élément particulièrement important. Le microbiote intestinal s'installe immédiatement après la naissance et évolue progressivement au cours de la vie de l'hôte. Chez les Mammifères, l'influence du microbiote vaginal de la mère est très importante dans la mise en place du microbiote chez le jeune. Chez les ovipares, l'influence des parents est moindre, mais après l'éclosion, les contacts physiques avec les parents facilitent également la mise en place d'un microbiote intestinal sain chez l'oisillon. Chez la poule d'élevage, les œufs sont conservés dans une écloserie et l'éclosion a lieu loin des parents. Le microbiote des jeunes poussins est alors principalement influencé par l'environnement d'élevage, la litière, la nourriture, le personnel qui les manipule, et toutes les surfaces avec lesquelles ils peuvent entrer en contact (Kers et al. 2018). La mise en place du microbiote se fait rapidement : au bout de 7 jours, la plupart des espèces sont présentes, mais elles mettront plusieurs semaines à voir leur abondance se stabiliser (Videnska et al. 2014; Khan et al. 2020) (**Figure A-6-B**). Le temps de stabilisation de ce microbiote ne fait pas consensus dans la littérature en particulier parmi les études sur le poulet de chair, mais toutes s'accordent au maintien de cette stabilité dans le temps. A l'échelle du phylum, le microbiote cæcal de poule pondeuse passe d'une dominance de Bacillota à un jeune âge, à une dominance équilibrée de Bacillota et de Bacteroidota à l'âge adulte (entre la 20^{e} et la 30^{e} semaine) (Videnska et al. 2014) et qui reste globalement assez stable ensuite.

Du fait de cette multitude de facteurs influençant la composition du microbiote, il est important de comparer les études provenant de contextes expérimentaux proches. Pour la poule, la littérature sur les études du microbiote intestinal est largement dominée ces dernières années, par des études sur les poulets de chair (971 contre 104 références bibliographiques disponibles sur PubMed entre 2011 et 2024 pour les poulets de chair ou les poules pondeuses). Or, même si beaucoup de genres de micro-organismes sont retrouvés en commun entre les poulets de chair et les poules pondeuses, ces deux catégories de poules diffèrent sur de multiples aspects. Les souches spécifiques aux poulets de chair sont sélectionnées sur des caractères de production différents (vitesse de croissance, gain de poids).

Elles ont donc des besoins physiologiques et nutritifs différents ce qui aboutit à des régimes alimentaires également différents. Par ailleurs, la durée d'élevage d'un poulet de chair se compte en jours (35 pour les poulets à croissance rapide et 100 jours pour le label rouge à croissance lente) alors que celle des poules pondeuses se compte en semaines (environ jusqu'à 90 à 100 semaines). Ces différences de génotypes, d'âge, et d'environnements sont autant de facteurs expliquant les différences de composition microbienne observées entre les études sur les poulets de chair et celles sur les poules pondeuses. Cela illustre également la difficulté de comparaison des études et la nécessité d'étudier plus précisément le microbiote spécifiquement des poules pondeuses dans des contextes variés d'âges, de génotypes et de conditions d'élevage.

Ce contexte montre la grande sensibilité du microbiote à une multitude de facteurs biologiques ou environnementaux, mais au-delà de ces facteurs, les différentes méthodes d'analyses expérimentales et bioinformatiques peuvent influencer sensiblement les résultats.

A.4. Evolution des méthodes d'analyses permettant d'explorer les écosystèmes microbiens

L'analyse des microbiotes a été rendue possible par des découvertes majeures en microbiologie et par des innovations techniques. Les premières bactéries ont été découvertes et observées au microscope par Antoni van Leeuwenhoek en 1676. Au cours de la seconde moitié du XIXe siècle, les premières cultures bactériennes ont permis à des scientifiques tels que les Français Casimir Davaine, Henri Toussaint, Louis Pasteur, et à l'Allemand Robert Koch d'isoler et de caractériser les bactéries responsables de maladies telles que le choléra des poules, l'anthrax du mouton, la tuberculose, ou le choléra chez l'homme. À cette époque, les micro-organismes sont essentiellement vus comme organismes nocifs. Robert Koch confirme les postulats de la pathogénicité en 1881 (Koch 1881). Par ailleurs, Louis Pasteur décrira également le processus de fermentation, notamment la fermentation lactique affectant les produits laitiers, et la fermentation alcoolique due aux levures et affectant la production d'alcool de betterave. Faisant suite aux travaux de Lazzaro Spallanzani du XVIII^{e} siècle, Louis Pasteur en 1860, démontrera la présence de micro-organismes vivants dans l'air, réfutant ainsi la théorie de la génération spontanée selon laquelle la vie pouvait naître de la matière inanimée. La fin du XIX^{e} siècle voit également se développer les travaux en écologie microbienne mettant en évidence le caractère ubiquitaire et également bénéfique des micro-organismes. Sergei Winogradsky a prouvé la présence de micro-organismes dans une diversité importante d'environnements, y compris sur des substrats inorganiques. Martinus Beijerinck a, quant à lui, découvert le principe de la fixation de l'azote dans le sol par des bactéries, un processus essentiel pour la nutrition de certaines plantes. Mais tous

ces travaux reposent sur la capacité à créer un environnement de culture, en particulier un milieu dédié pour chaque type de micro-organisme pour pouvoir les caractériser. La découverte de l'ADN, sa caractérisation en 1953 par James Watson, Francis Crick, Maurice Wilkins, et Rosalind Franklin révolutionne la science, tout comme les progrès en biologie moléculaire qui aboutissent au développement de techniques de séquençage de cette molécule. L'acquisition de cette technique et de ses évolutions – Sanger en 1977, séquenceurs de nouvelle génération (Roche 454, ABI SOLiD, Solexa, Illumina) au début des années 2000, séquenceurs de troisième génération (PacBio, Nanopore) à partir de 2010 (Ebertz 2020) – bouleversent les connaissances en permettant la caractérisation d'écosystèmes microbiens indépendamment de leur capacité à être cultivés ou pas.

Ces nouvelles analyses de microbiotes ont en particulier révélé la faible proportion de micro-organismes jusque-là cultivables, environ 30% pour les espèces du microbiote humain (Almeida et al. 2021), moins de 20% chez la poule (Oakley et al. 2014), et encore moins pour des écosystèmes environnementaux comme la mer (6%) (M. Wang et al. 2020). Même si des progrès importants sont fait en culturomique (Diakite, Dubourg, and Raoult 2021), les protocoles de séquençage ont l'avantage indéniable de proposer une solution nettement plus exhaustive de la caractérisation des microbiotes. Ces nouvelles techniques ont également eu un impact sur la caractérisation des espèces et de leur taxonomie, en particulier celle des procaryotes. Historiquement, les procaryotes étaient classés selon différentes caractéristiques phénotypiques (croissance, morphologie, composition chimique) ou encore selon leur potentiel pathogène (Schleifer 2009). Avec les progrès de la biologie moléculaire et des connaissances sur l'ADN, le concept d'espèce est redéfini en fonction du taux d'hybridation ADN/ADN et/ou de la température de fusion (température à laquelle l'ADN passe de la forme simple brin à la forme double brin). Ainsi, au sein d'une même espèce, le taux d'hybridation entre ADN de deux souches d'une même espèce doit être de plus de 70%, et leur température de fusion doit être différente de moins de 5%. Avec l'avènement du séquençage, la phylogénétique se développe et permet sur la base de similarités de séquences, non seulement de séparer les espèces entre elles mais également de mesurer leur relation de parenté. C'est notamment grâce à ces avancées technologiques que Carl Woese et George Fox utilisent la séquence de l'ARN ribosomal 16S (ARNr 16S, constituant la petite sous-unité des ribosomes chez les procaryotes) pour décrire, en 1977, les archéobactéries comme étant différentes des bactéries (Woese and Fox 1977). Il proposera, en 1990, de créer un troisième règne spécifique des archées (Archaea), au côté des bactéries (Bacteria) et des eucaryotes (Eucarya ou Eukaryota) (Woese, Kandler, and Wheelis 1990). L'ARNr 16S s'est ensuite imposé comme le gène marqueur de référence chez les procaryotes. Avec l'évolution des techniques de séquençage permettant une augmentation du débit, une diminution du taux d'erreur et du coût financier, cette approche phylogénétique basée sur la similarité de séquence, s'applique dorénavant à un ensemble

de gènes marqueurs voire aux génomes entiers. Au-delà de l'utilisation de ces séquences en phylogénie, ces techniques accélèrent la recherche sur les microbiotes avec des projets d'envergures, notamment chez l'homme ou dédiés à différents environnements (Human Microbiome Projects (Human Microbiome Project Consortium 2012; 2014), EarthMicrobiome Project (J. A. Gilbert, Jansson, and Knight 2014), TerraGenome (Vogel et al. 2009) ou Tara Ocean (Pesant et al. 2015)). Ces projets ont pour but de caractériser les microbiotes du point de vue de leur richesse (quelles sont les micro-organismes présents ?) ; leur diversité (en quelle abondance ?) ; leur dissimilarité (sont-ils conservés entre individus/échantillons ? Sont-ils influencés par des caractéristiques de leur hôte ou leur environnement ?) ; leur fonction (qu'apportent-ils à leur hôte ou à leur environnement ?).

Ces notions de richesse et de diversité viennent du domaine de l'écologie, et se divisent en diversités alpha et bêta (**Figure A-7**). La diversité alpha évalue la diversité d'un unique microbiote tandis que la diversité bêta reflète la diversité des microbiotes qui peuvent composer un écosystème. Pour chacune de ces diversités, il existe plusieurs indices qui apportent chacun des nuances différentes. Pour l'alpha diversité, on peut citer notamment les indices de richesse et les indices de diversité. Les premiers, par exemple la richesse observée, ou l'indice de Chao1 qui estime les espèces non détectées, comptabilisent le nombre d'espèces différentes. Plus il y a d'espèces, plus le microbiote est riche. Les seconds, par exemple l'indice de Shannon ou l'inverse de l'indice de Simpson, évaluent l'homogénéité des abondances de chacune de ces espèces. A richesse équivalente, plus les abondances sont équilibrées et plus le microbiote est divers (**Figure A-7**, échantillon S1 versus S2).

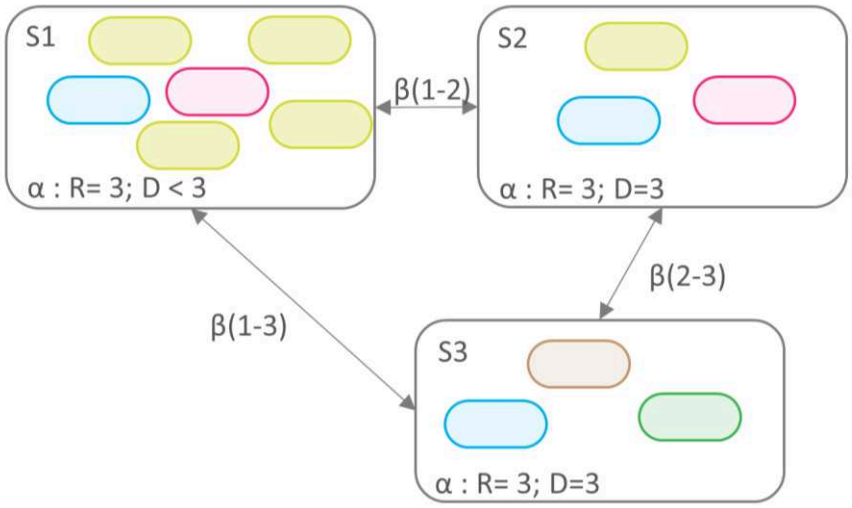


Figure A-7 : Schéma des indices de diversité alpha et bêta.
Chaque rectangle gris représente un échantillon et chaque élément coloré une cellule microbienne dont le type se différencie par la couleur. Les indices de diversité alpha sont calculés pour chaque échantillon (R : richesse, D : diversité), et les indices de diversité bêta se calculent entre paire d'échantillons (représenté par la longueur des flèches).

La bêta diversité compare deux échantillons, et l'ensemble des comparaisons des échantillons deux à deux permet de construire une matrice de distances, que l'on appelle également matrice de dissimilarité. Il existe une multitude de méthodes pour calculer ces distances. Parmi les plus utilisées, on retrouve des mesures qualitative (distance de Jaccard), quantitative (distance de Bray Curtis), incluant ou non des notions de distance phylogénétique (distance Unifrac éventuellement pondérée par l'abondance). Entre deux échantillons, plus il y a d'espèces qui sont spécifiques ou en surabondance et/ou éloignées phylogénétiquement, plus les microbiotes sont distants l'un de l'autre (**Figure A-7**, échantillon S1 versus S3), et *in fine* plus l'écosystème (et tous les échantillons possibles qui le composent) est divers. Ces indices, définis dans le cadre des études en écologie, s'appliquent en premier lieu à toutes entités taxonomiques (souches, espèces, genres, …), mais ils peuvent également, pour une partie d'entre eux, s'appliquer à toutes entités quantifiables notamment aux gènes ou aux fonctions portés par le microbiote.

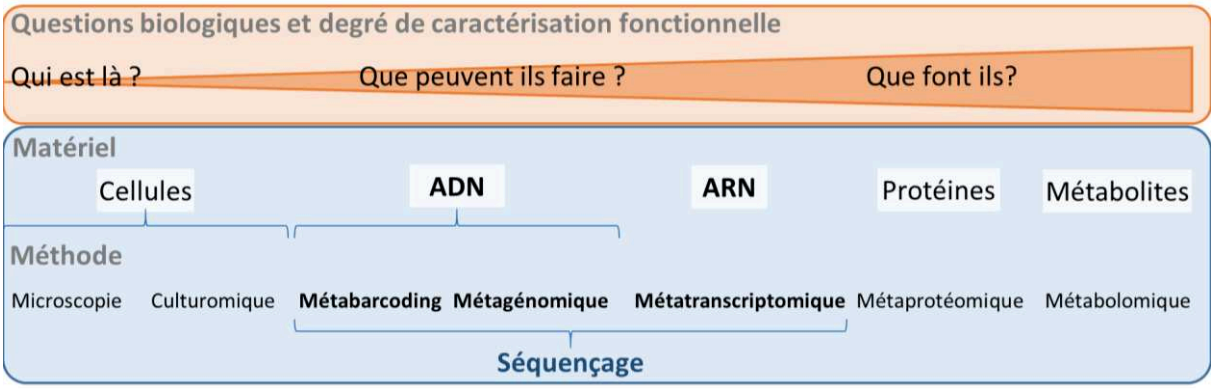


Figure A-8 : Différents protocoles d'analyses du microbiote.
Figure modifiée à partir de Berg et al. 2020.

A ce jour, plusieurs protocoles expérimentaux existent et permettent plus ou moins précisément de caractériser les micro-organismes présents (par leur taxonomie et leur abondance) et leurs fonctions (potentielles ou actives) (**Figure A-8**). Parmi les protocoles impliquant du séquençage et donc ne nécessitant pas de culture préliminaire, il y a la méthode de métabarcoding ciblant un gène marqueur amplifié (ou séquençage amplicon), la métagénomique (ou « *shotgun sequencing* ») séquençant l'ensemble de l'ADN de l'ensemble des organismes, et la métatranscriptomique ciblant l'ensemble des ARN (en particulier les ARN messagers). Ces trois protocoles permettent d'explorer à la fois la composition microbienne et les (potentielles) fonctions portées par les micro-organismes. En plus des méthodes basées sur le séquençage, on retrouve les méthodes basées sur de la culture qui bien que moins exhaustives permettent d'aller plus loin dans la caractérisation individuelle des micro-organismes et de valider leurs fonctions ; ainsi que la métaprotéomique et la métabolomique.

Ces dernières, potentiellement indépendantes d'une culture préliminaire, permettent d'explorer de façon exhaustive les productions du microbiote. Dans la mesure où notre étude n'inclut pas ces dernières méthodes, nous nous focaliserons ensuite seulement sur la description des méthodes basées sur le séquençage i.e. le métabarcoding, la métagénomique et la métatranscriptomique.

A.5. Méthodes contemporaines basées sur le séquençage

A l'image des évolutions faites dans le domaine de la caractérisation des taxonomies, les protocoles de séquençage ont vu d'abord se populariser la technique de métabarcoding en particulier sur l'ARNr 16S toujours très utilisé aujourd'hui, puis la métagénomique qui voit son utilisation progresser largement et enfin la métatranscriptomique qui est encore aujourd'hui un protocole très novateur (**Figure A-9**).

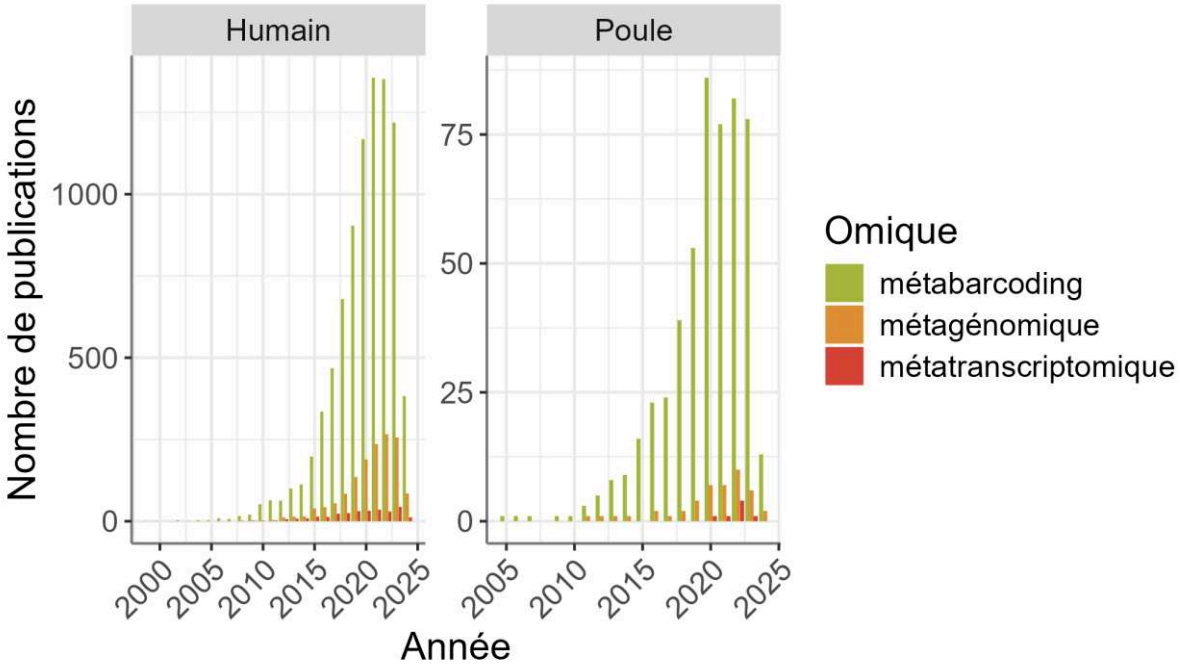


Figure A-9: Nombre de publications pour l'étude du microbiote intestinal chez l'homme ou la poule en fonction du protocole de séquençage utilisé.
Interrogation Pubmed avec les termes `chicken` ou `human` et `((gut microbiota) OR (gut microbiome))` et pour le métabarcoding : `(metabarcoding) OR (amplicon) OR (metataxonomics) OR (16S sequencing)` ; pour la métagénomique : `(shotgun metagenomics) OR (whole metagenome sequencing) OR (metagenome-assembled genome) or (MAG)` ; et pour la métatranscriptomique : `(metatranscriptomic) OR (metatranscriptome).`

La première publication dédiée à l'analyse du microbiote humain mentionnant le séquençage par métabarcoding date de 1999 (Suau et al. 1999) et de 2008 pour la première étude mentionnant l'utilisation de la métagénomique (Tito et al. 2008). Chez la poule, ces deux techniques ont été

mentionnées pour la première fois en 2011 (Torok et al. 2011; Danzeisen et al. 2011), mais le nombre d'étude pour chacun de ces techniques a évolué différemment au cours du temps. En effet, bien que le nombre de publications utilisant le séquençage complet de l'ADN, les études utilisant le métabarcoding dominent encore aujourd'hui largement la littérature et représentent plus de 80% et 91% des publications pour chacun de ces écosystèmes microbiens. Enfin, les premières études du microbiote intestinal humain utilisant la métatranscriptomique sont publiées à partir de 2011 (Gosalbes et al. 2011) et il faut attendre 2020 pour des études chez la poule (Y. Wang et al. 2020). Les paragraphes suivants détaillent le principe du protocole et d'analyse de chacune de ces méthodes.

Au-delà du fait que ces protocoles ne répondent pas toujours aux mêmes questions, leur variation de popularité d'utilisation peut s'expliquer par des causes techniques, humaines et financières. Ainsi le protocole expérimental est plus complexe pour la métatranscriptomique. Les outils d'analyses sont moins développés pour la métagénomique et la métatranscriptomique et ces protocoles nécessitent des ressources informatiques de calculs et de stockage plus importantes. Enfin le métabarcoding est nettement plus abordable que les deux autres protocoles. Sur notre projet, en 2022, le coût du séquençage (hors extraction) d'un échantillon en métabarcoding était de l'ordre de 27€, alors que celui de la métagénomique était de 185€ et celui de la métatranscriptomique, en tenant compte de la ribodéplétion, était de 257€.

A.5.1. Le métabarcoding : séquençage d'un gène marqueur

a. Principe du protocole expérimental

Le principe du protocole expérimental et d'analyse du métabarcoding est simple (**Figure A-10**). L'idée est de ne séquencer qu'une seule région des génomes de tous les organismes présents dans le milieu étudié. Cette région doit être ubiquitaire (ou présente dans le plus grand nombre possible de génomes) mais suffisamment différente pour discriminer les génomes en fonction des groupes taxonomiques auxquels ils appartiennent. En plus d'être ubiquitaire et discriminant, le gène ou la région doit être amplifiable. Il doit donc contenir des régions conservées entre les différentes génomes permettant de définir des amorces utilisables lors d'une PCR (« *Polymerase Chain Reaction* ») pour amplifier la région cible et générer ainsi des amplicons qui seront ensuite séquencés. Ces amplicons doivent par ailleurs, respecter des contraintes de taille imposées par les techniques de séquençage. Le séquençage Roche 454 populaire dans les années 2010, permettait de séquencer des lectures simples de tailles variables allant en moyenne de 400 à 600 pb. Le séquençage Illumina en particulier dans sa version MiSeq, est passé de lectures pairées allant de 2x150 pb à 2x300 pb. Ces librairies de séquençage permettent donc l'utilisation de régions de gènes marqueurs de tailles inférieures à 600 pb. Avec la 3^{e} génération de

séquenceur, les lectures atteignent entre 15 et 20 kb pour PacBio et 10 à 100kb pour Nanopore, levant ainsi cette contrainte de taille. Ce type de séquençage « *long-read* » est toutefois encore peu utilisé (à ce jour, 39 publications sur des écosystèmes intestinaux). Enfin pour répondre au but premier du métabarcoding, *i.e.* l'identification des communautés microbiennes présentes, le gène cible doit être présent et annoté taxonomiquement dans les bases de données. Actuellement, ce sont les gènes liés à la machinerie cellulaire, transcription et traduction qui sont le plus utilisés (Schleifer 2009; Berg et al. 2020; Casey et al. 2021) avec comme « étalon d'or » pour les procaryotes le gène de l'ARNr 16S et pour les eucaryotes celui de l'ARNr 18S.

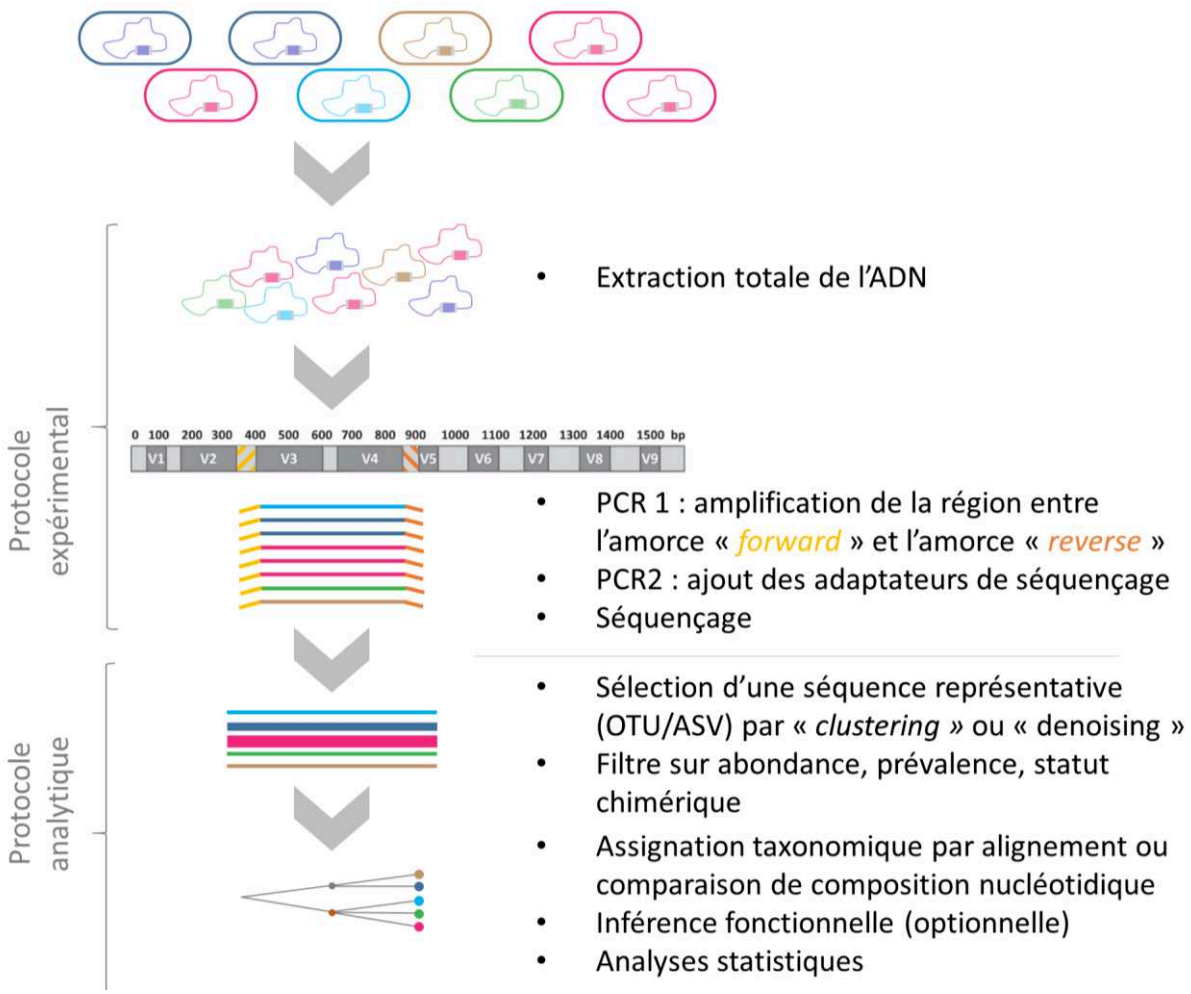


Figure A-10: Schéma du protocole expérimental et analytique du métabarcoding.
Exemple de la sélection de la région variable V3-V4 du gène de l'ARNr 16S comme marqueur cible. Le schéma de l'alternance des régions conservées (gris claire) et variables (gris foncé) de l'ARNr 16S provient de McAllister et al. 2018.

b. Principe de l'analyse de séquences

L'analyse des séquences est ensuite relativement intuitive, les séquences représentant comme un code barre de l'identité de chaque groupe taxonomique (**Figure A-10**). Elle comprend en général une étape de nettoyage des séquences dite de « *preprocessing* », une étape de sélection des séquences représentatives des différents groupes taxonomiques, éventuellement une étape de filtre, et finalement une étape d'annotation taxonomique des séquences représentatives. Il existe un grand nombre d'outils et de stratégies pour chacune de ces étapes qui ont notamment été répertoriés et décrits dans une revue dont je suis co-auteure et publiée lors de cette thèse (Hakimzadeh et al. 2023, également disponible en annexe **J.1**).

L'objectif général du *preprocessing* est de nettoyer et de préparer les séquences brutes issues d'amplifications par PCR et du séquençage.

L'étape suivante est l'identification des séquences représentatives des différents organismes de l'écosystème étudié. Elle représente le cœur de l'analyse de métabarcoding. C'est majoritairement cette étape qui a évolué ces dernières années et qui différencie encore aujourd'hui les stratégies d'analyses. Une OTU (« *Operational Taxonomic* Unit ») définit à l'origine un ensemble d'organismes similaires d'un point de vue d'un certain nombre de caractères (Sokal 1963). A l'échelle de l'analyse de séquences, une OTU représente donc un groupe de séquences similaires. Pour rappel, l'identification d'une espèce procaryote a été initialement définie notamment en fonction de critères de taux d'hybridation ADN/ADN et de température de fusion. Avec le dépôt de plus en plus de séquences du gène de l'ARNr 16S, il a été possible d'explorer le pourcentage de similarité entre séquences d'espèces (ou de rang taxonomique supérieur) répertoriées comme différentes sur ces précédents critères. De cette analyse est ressorti le fait que pour atteindre une hybridation d'au moins 70% et donc appartenir à la même espèce, les séquences des gènes de l'ARNr 16S devaient partager au moins 97% d'identité (Stackebrandt and Goebel 1994). Ce seuil a été réévalué à 99% d'identité en 2006 (Stackebrandt and Ebers 2006; Rosselló-Móra and Amann 2015; Edgar 2018). Les premières stratégies d'analyse de séquences (par exemple mothur ou uparse) ont alors repris ces seuils pour paramétrer des outils de « *clustering* » de séquences, l'idée étant de regrouper les séquences partageant au moins 97% d'identité entre elles (et 99% plus tard) (Schloss et al. 2009; Edgar 2013). En parallèle d'autres méthodes de clustering sont apparues telle que swarm dont le but n'est pas de regrouper les amplicons selon un seuil global comme décrit précédemment mais repose sur un petit seuil de liaison local, représentant le nombre maximum de différences entre deux amplicons (Mahé et al. 2014). Ces nouveaux algorithmes tendent davantage à prendre en compte les procédés techniques d'obtention des amplicons ainsi que les histoires évolutives des espèces. A partir de 2016, de nouvelles méthodes (par exemple DADA2 ou Deblur) se développent dont l'objectif est de conserver la diversité génétique

des amplicons sans se focaliser sur la séparation d'un rang taxonomique donné (Callahan et al. 2016; Amir et al. 2017). Ces outils tirent profit des connaissances acquises sur les technologies de séquençage pour appliquer un modèle d'erreur prédéfini en fonction du type de séquenceur, et corriger les nucléotides qui seraient dus à des erreurs de séquençage plutôt qu'à de vraies mutations. Ces méthodes de « *denoising* » produisent ce que l'on appelle majoritairement des ASV pour « *Amplicon Sequence Variant* ».

Une grande partie des stratégies d'analyse propose ensuite de filtrer les séquences représentatives artéfactuelles des OTU/ASV, en particulier sur leur statut chimérique ou non, et sur leur niveau de représentativité dans l'échantillonnage. En effet, l'un des biais techniques notable du métabarcoding est la création de séquences chimériques dues à l'utilisation successive de deux amplifications par PCR. Par ailleurs, il est considéré que les séquences en très faible abondance sont probablement des séquences artéfactuelles (Bokulich et al. 2013). En effet, puisque ces séquences sont issues d'amplifications par PCR, il est très peu probable qu'une séquence « vraie » reste en un seul exemplaire (ou en un très faible nombre) dans le jeu de séquences générées. Ces séquences de trop faible abondance seraient plutôt le fruit d'accumulation d'erreurs pendant les amplifications, et le séquençage.

Une fois les séquences regroupées et filtrées, une dernière étape consiste à leur assigner une affiliation taxonomique par comparaison à une base de référence du même marqueur (détaillée au paragraphe suivant **A.5.1.c**). Pour cela il existe différentes méthodes, j'en présenterai deux ici. La première est basée sur des outils d'alignement de séquences (par exemple Blast+), alors que la seconde compare les compositions nucléotidiques (par exemple RDP classifier) (Q. Wang et al. 2007; Camacho et al. 2009). Les outils d'alignements alignent les séquences des OTU/ASV contre les séquences contenues dans une banque de référence de séquences annotées. Des métriques caractérisant les alignements sont produites, en particulier le pourcentage d'identité et de couverture de la séquence d'intérêt vis-à-vis de la séquence de référence. Il est alors possible de mesurer la distance entre ces deux séquences. Plus la distance est petite, plus la séquence d'intérêt est proche taxonomiquement de la séquence de référence. Les outils de composition nucléotidique comparent la composition en mot de quelques nucléotides de chaque séquence d'OTU/ASV avec l'ensemble des séquences de référence. Pour confirmer l'attribution de la taxonomie de la référence la plus proche, un score de « *bootstrap*» est calculé pour chaque rang taxonomique. Ce score correspond à la fréquence d'apparition du rang parmi les reclassifications d'un sous ensemble aléatoire des mots de la séquence d'OTU/ASV. Ainsi, alors que les méthodes d'alignement fournissent des mesures directes de similarité entre deux séquences, la mesure de « *bootstrap* » indique un score de confiance de l'annotation retournée vis-à-vis de l'ensemble des autres possibilités. De fait, si la base de séquences de référence utilisée est très

éloignée de l'écosystème étudié, *i.e.* non exhaustive vis-à-vis des organismes présents dans cet écosystème, les méthodes d'alignement pourraient retourner des scores de qualité faible alors que les méthodes de composition pourraient retourner des scores de confiance forts. Au-delà de la méthode, c'est le choix de la base de données de référence qui va influencer la qualité de l'assignation taxonomique (Hakimzadeh et al. 2023).

Grâce à l'accumulation des connaissances dans les bases de données, en particulier de génomes complets, et bien que ce ne soit pas le but initial du séquençage métabarcoding, des outils (par exemple tax4fun2 ou PICRUSt2) permettant l'analyse des fonctions portées par le microbiote ont été développés spécifiquement pour ce type de données (Wemheuer et al. 2020; Douglas et al. 2020). Pour cela, ils font appel à de l'inférence fonctionnelle, c'est-à-dire une prédiction des fonctions à partir du fragment d'ADN séquencé et non de l'ensemble des gènes directement porteurs de ces fonctions. Ces outils reposent sur des bases de données fonctionnelles pré-calculées. Elles sont créées à partir des génomes complets annotés taxonomiquement et fonctionnellement. Les séquences de gènes marqueurs (majoritairement 16S) sont extraites de ces génomes et forment ainsi la base de référence contre laquelle les séquences des OTU/ASV sont comparées. Cette référence peut être représentée simplement sous la forme de séquences ou bien d'un arbre phylogénétique. La comparaison entre les séquences des OTU/ASV et la référence se mesurera *via* un pourcentage de similarité ou *via* le calcul d'une distance phylogénétique. Les fonctions connues des génomes complets sont ensuite réattribuées aux OTU/ASV *via* leur proximité avec une séquence du gène marqueur de référence.

c. Bases de références taxonomiques

La taxonomie permet d'identifier des groupes d'organismes en fonction de noms scientifiques (les taxons) et d'une classification hiérarchisée. Alors que la nomenclature des taxons et leur officialisation sont régies par différents codes spécifiques de différents groupes d'organismes – l'ICNP (« International Code of Nomenclature of Prokaryotes ») pour les procaryotes (bactéries et archées) ; l'ICVCN (« *International Code of Virus Classification and Nomenclature* ») pour les virus ; l'ICNafp (« International Code of Nomenclature for algae, fungi, and plants ») pour les algues, les champignons et les plantes ; et l'ICZN (« International Code of Zoological Nomenclature ») pour les animaux – il n'existe pas de consensus sur leur classification hiérarchique. La classification évolue donc en fonction de la mise à jour des codes et l'officialisation des noms des taxons, mais également selon les bases de références qui incluent potentiellement des taxonomies encore non reconnues officiellement et utilisent des méthodes de classification différentes.

Il existe une multitudes bases de références. Elles peuvent être généralistes ou spécifiques d'un gène marqueur, d'un clade ou encore d'un milieu. Parmi les bases de données les plus communément utilisées, on trouve les bases généralistes GenBank ou RefSeq, desquelles il faut extraire les séquences du gène marqueur cible qui sont associées à la taxonomie du NCBI (Schoch et al. 2020). La classification taxonomique du NCBI a initialement été construite en intégrant les classifications phylogénétiques des trois principales bases de données de séquences nucléotidiques (GenBank, EMBL et DDBJ). Elle inclut également des données de la littérature et de bases spécialisées dans certains groupes taxonomiques pour l'expertiser et l'abonder (NCBI Staff, n.d.). Elle correspond désormais à la taxonomie standard pour les bases du NCBI, d'EMBL et de DDBJ. Un comité d'experts est chargé de faire évoluer cette taxonomie en fonction de l'ajout de nouvelles séquences représentant de nouveaux taxons et en reportant les mises à jour des différentes nomenclatures taxonomiques. Les dernières grandes modifications comprennent par exemple, la prise en compte de la nouvelle nomenclature des phyla chez les procaryotes (suffixes –ota) ou la mise à jour des espèces de virus grippaux en noms binomiaux (*i.e.* en deux termes) (NCBI Staff 2022; 2023).

D'autres bases de données largement utilisées sont dédiées aux séquences ribosomales. La base de données SILVA contient des séquences provenant d'EMBL des gènes codant les petites (16S et 18S) et grandes sous-unités (23S et 28S) des ribosomes et ce pour les trois domaines du vivant, Bacteria, Archaea, Eukarya (Quast et al. 2013; Yilmaz et al. 2014). Pour chaque sous-unité, un arbre phylogénétique guide est créé tirant profit d'autres projets de grande expertise comme le « *All-Species Living Tree Project* » (LTP) (Ludwig et al. 2021). L'annotation taxonomique est ensuite expertisée en intégrant les taxonomies de différentes sources, notamment de l'index « *Bergey's Taxonomic Outlines* » (Garrity, Bell, and Lilburn 2004) et de la «*List of Prokaryotic Names with Standing in Nomenclature* » (LPSN, qui intègre une grande diversité de sources) (Parte 2018), et des analyses phylogénétiques d'autres bases de données (comme la « *Genome Taxonomy Database* »(GTDB) (Parks et al. 2022) ou UniEuk (Berney et al. 2017)). L'ensemble des séquences sont ensuite affiliées en fonction de cet arbre. La base du « *Ribosomal Database Project* » (RDP (Cole et al. 2014)) est dédiée aux gènes ribosomaux (codant les sous-unités 16S pour les procaryotes et 28S pour les champignons). Pour chaque gène, RDP produit un alignement multiple basé sur une référence expertisée et annotée. La classification taxonomique s'appuie ensuite, tout comme SILVA, sur l'index « *Bergey's Taxonomic Outlines* » et sur le LPSN. Pour les eucaryotes et en particulier les champignons, la base de données UNITE fait référence. Elle distribue des séquences majoritairement pleine longueur des régions « *Internal Transcribed Spacer* » (ITS) positionnées entre les gènes de la petite et la grande sous-unité ribosomale (Abarenkov et al. 2024). Les séquences sont traitées par une suite de « *clustering* » à différents niveaux de similarité pour produire de « *Species Hypotheses* » (SH). Celles-ci sont annotées

à partir de l'intégration de taxonomies provenant de différentes sources d'annotations taxonomiques, certaines étant spécifiques des champignons (« *Outline of Fungi* » (Wijayawardene et al. 2022) ou MycoBank (Robert et al. 2013)) et d'autre généraliste des eucaryotes (GBIF (GBIF Secretariat 2023)).

Enfin, il existe de nombreuses bases de références spécialisées d'un milieu ou d'un clade plus précis. Par exemple, MiDAS (Dueholm et al. 2024) est spécifique des écosystèmes des boues activées et des digesteurs anaérobies. Elle propose des séquences d'ARNr 16S affiliées taxonomiquement grâce à SILVA. DAIRYdb (Meola et al. 2019), également une base de données de séquences d'ARNr 16S est spécifique des organismes identifiés dans les produits laitiers. Les taxonomies sont basées sur la taxonomie SILVA mais expertisées et nettoyées manuellement. Au contraire, MIDORI (Leray, Knowlton, and Machida 2022) et Diat.barcode (Rimet et al. 2019) sont des références spécifiques de clades particuliers, les animaux (Metazoa) et les diatomées (Bacillariophyta). Les séquences et leur taxonomie proviennent en partie du NCBI.

d. Biais et difficultés liés à l'analyse de séquences métabarcoding

Bien que le métabarcoding présente de nombreux avantages – facilité expérimentale, faible coût financier, analyse de plus en plus standardisée, bases de données de référence de plus en plus complètes et une expertise globale de la communauté scientifique forte *via* sa popularité d'utilisation – il présente également des inconvénients et des difficultés, principalement liés au choix du gène marqueur et au protocole expérimental basé sur de l'amplification PCR (**Figure A-11**). Ces inconvénients dus à des raisons biologiques ou techniques sont largement documentés notamment pour proposer des solutions pour les atténuer (Bonk et al. 2018; Moinard et al. 2023).

Le principe du métabarcoding repose sur le choix d'un gène marqueur qui serait universel et discriminant. Or l'universalité et la capacité discriminante parfaite n'existent pas. Par exemple, le gène de l'ARNr 16S, gène marqueur par excellence des études sur le microbiote intestinal (Deusch et al. 2015), ne permet pas la détection des eucaryotes. A un niveau plus précis, les régions conservées, dans lesquelles les amorces permettant l'amplification de la région cible sont définies, ne sont pas strictement conservées et certains organismes porteurs du gène marqueur peuvent ne pas être capturés. Pour atténuer ce biais de représentativité, il est possible d'utiliser de bases dégénérées permettant quelques différences dans la séquence des amorces.

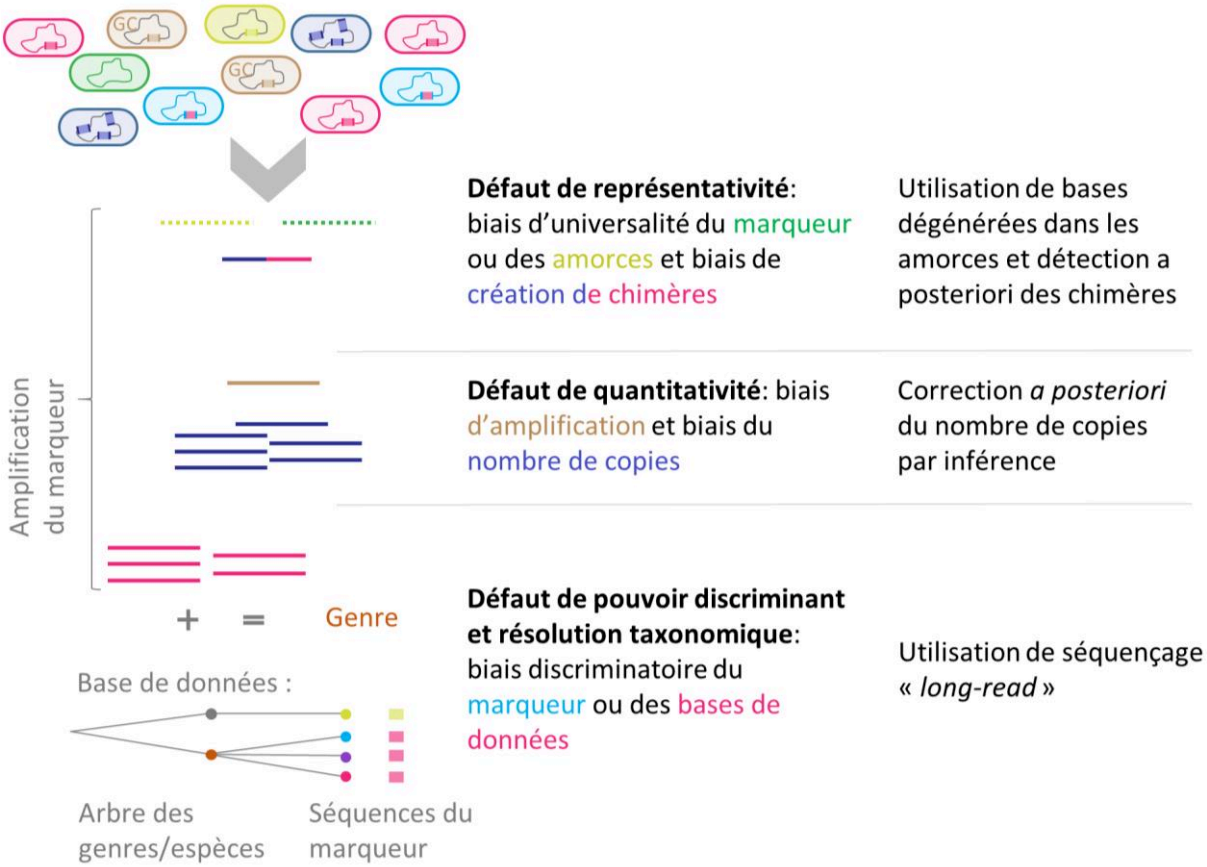


Figure A-11: Représentation des biais naturels et techniques du métabarcoding et de solutions correctives.

Le pouvoir discriminant, quant à lui, est lié au pourcentage de similarité du gène marqueur entre différents organismes (notamment entre différentes espèces). Or le gène de l'ARNr 16S présente un fort degré de conservation (Větrovský and Baldrian 2013; Rosselló-Móra and Amann 2015), qui a amené à augmenter le seuil recommandé pour différencier deux espèces à 99%. Ce seuil équivaut à seulement 15 différences sur une séquence d'environ 1 500 nucléotides. Ainsi, on peut aisément imaginer que la taille de la région cible va influencer le nombre de différences observables permettant de différencier les espèces présentes. Or, à cause de contraintes de taille de fragment séquençable sur les séquenceurs Illumina couramment utilisés actuellement, la région ciblée est souvent réduite à une sous-région du gène cible. Pour l'ARNr 16S, selon la région ciblée, le pouvoir discriminant des espèces est variable et systématiquement moins important que celui du gène complet (Johnson et al. 2019). Le séquençage « long-read » rendu possible par les séquenceurs de 3^{e} génération permet d'éliminer cette contrainte et limiter ce défaut de pouvoir discriminant (Wagner et al. 2016).

Enfin, pour être quantitative, l'analyse de métabarcoding doit s'appuyer sur un gène marqueur présent en simple copie dans chaque génome. Or, en ce qui concerne l'ARNr 16S, il a été montré en 2013 que seul 15% des espèces bactériennes n'avaient qu'une seule copie de ce gène (Větrovský and Baldrian

2013). En 2022, selon la version 5.8 de la base de données rrnDB (Stoddard et al. 2015), les bactéries avaient en moyenne 1,99 copies et jusqu'à 21, et les archées 1,1 et jusqu'à 5 copies. Ce biais peut être atténué lors de l'analyse bioinformatique des séquences en inférant le nombre de copies à partir des connaissances accumulées sur les assemblages complets des génomes procaryotes. A noter que l'analyse de métabarcoding, même sur le gène de l'ARNr 16S, n'empêche pas une analyse quantitative entre échantillon. Pour pallier ces biais biologiques dont certains sont spécifiques aux gènes ribosomaux, d'autres gènes marqueurs peuvent être utilisés. En particulier, le gène rpoB, codant la sous-unité β de l'ARN polymérase possède également des régions variables et conservées, est présent en simple copie, et a montré des résultats prometteurs (Ogier et al. 2019). Toutefois ces régions conservées sont trop variables pour définir des amorces capables de capter l'ensemble de la diversité bactérienne (Hassler et al. 2022), et il n'existe pas à ce jour de bases de données dédiées à ce marqueur.

Le caractère non quantitatif du métabarcoding est également dû à l'utilisation de deux amplifications par PCR qui apportent un certain nombre de biais techniques. En effet, l'amplification par PCR ne se fait pas de manière homogène ou en respect des abondances des fragments d'ADN initiaux : i) les séquences riches en GC sont moins amplifiées (Bonk et al. 2018), ii) les fragments plus courts sont mieux représentés ce qui biaise la quantification des ITS qui ont des tailles variables selon les organismes (De Filippis et al. 2017), iii) les cycles consécutifs d'amplification amènent à la formation de séquences chimériques, qui ont pour conséquence d'augmenter la richesse observée. Pour ce dernier biais, il est possible de limiter son effet en adoptant un protocole à une seule amplification (Rausch et al. 2019), en limitant le nombre de cycles d'amplification et grâce à une détection *a posteriori* lors de l'analyse des séquences et de leurs abondances. Enfin, comme vu précédemment, l'identification taxonomique des séquences sera biaisée par l'exhaustivité de la banque de référence utilisée.

Au-delà des solutions listées ci-dessus pour pallier un certain nombre d'inconvénients du métabarcoding (notamment réalisé sur l'ARNr 16S), la caractérisation d'un écosystème microbien peut se faire *via* d'autres méthodes comme la métagénomique complète. En effet celle-ci ne reposant pas sur une amplification PCR, elle ne souffre pas de ces mêmes biais techniques. Par ailleurs, en séquençant l'intégralité des gènes et des génomes, elle est moins sensible au défaut de pouvoir discriminant et permet une mesure directe des fonctions. Du fait notamment de son coût financier et de sa complexité d'analyse, la métagénomique complète ne remplace pas pour autant aujourd'hui le métabarcoding.

A.5.2. La métagénomique : séquençage complet des génomes

L'idée de cette méthode est d'être exhaustive dans la représentation des molécules d'ADN présentes. Comparée au métabarcoding, pour couvrir les génomes dans leur globalité, elle implique une plus grande profondeur de séquençage, d'autant plus pour identifier les micro-organismes en faible abondance, et donc un coût financier élevé. L'ADN total est extrait de l'échantillon, puis fragmenté aléatoirement avant d'être séquencé. Les séquenceurs utilisés actuellement sont généralement des séquenceurs Illumina qui génèrent des lectures de petite taille mais de grande qualité et à fort débit. Comme pour le métabarcoding, l'un des objectifs de cette méthode est de caractériser taxonomiquement l'ensemble des micro-organismes présents et de les quantifier. L'avantage majeur de la métagénomique est qu'elle permet en plus une annotation fonctionnelle directe du microbiote à partir de l'ensemble des gènes reconstitués et ne repose donc pas uniquement sur les génomes complets connus et annotés.

Il existe deux stratégies d'analyse de séquences de métagénomique qui remplissent deux objectifs différents. La première est l'analyse *de novo*, qui consiste en la reconstruction des gènes et des génomes à partir des lectures courtes, leur annotation et leur quantification. Son objectif est de construire une référence de gènes, de fonctions et de génomes présents dans un écosystème. Lorsqu'appliquée à de nombreux échantillons représentant le plus possible la diversité de l'écosystème, l'ensemble des gènes et des génomes constitue une référence communément appelée catalogue. La seconde se base sur une référence (un catalogue spécifique de l'écosystème ou des bases de données généralistes) et permet à partir d'un séquençage à moindre profondeur de quantifier les génomes et les fonctions connues. Cette méthode correspond à la métagénomique quantitative.

a. Analyse métagénomique de novo : principe de constitution d'un catalogue de référence

Un catalogue de référence en métagénomique est généralement constitué d'un ensemble de séquences de gènes et de génomes, obtenus à partir de séquences courtes qu'il faut assembler les unes aux autres (**Figure A-12**). Les techniques d'assemblage d'un génome unique reposent sur la reconstruction d'un graphe basé sur le chevauchement des lectures (par leur similarité de séquences notamment sur les extrémités), puis sur l'abondance de ces lectures pour déterminer les chemins possibles du graphe permettant de reconstruire un fragment d'ADN plus long et contigu, le contig. Ces contigs peuvent ensuite être associés et ordonnés en tenant compte de librairies de séquençage pairées qui vont faire le lien entre différents contigs. La métagénomique ajoute un niveau de complexité supérieur puisque l'écosystème étudié contient plusieurs génomes, potentiellement en très grand nombre, et avec des abondances souvent très déséquilibrées et majoritairement faibles

(Breitwieser, Lu, and Salzberg 2019). De plus ces génomes peuvent avoir des régions très conservées entre espèce et également être diversifiés entre souche d'une même espèce. Des outils d'assemblage dédiés ont donc été développés pour les séquences de métagénomiques (par exemple MEGAHIT ou MetaSPAdes (D. Li et al. 2016; Nurk et al. 2017)). Les contigs sont ensuite utilisés à deux fins : l'identification des gènes et des fonctions et la construction de génomes. Ils sont soumis à des outils d'annotation structurale (comme Prodigal) permettant l'identification des gènes par recherche de motifs et codons particuliers (Hyatt et al. 2010). Ces gènes sont ensuite annotés fonctionnellement par comparaison avec des bases de données de référence fonctionnelle (par exemple, alignement sur la base KEGG (Kanehisa et al. 2016), ou recherche d'orthologues sur la base intégrée eggNOG (Huerta-Cepas et al. 2019)). Par ailleurs, des outils de regroupement (« *binning* », par exemple MetaBAT2, MaxBin2) permettent d'associer les contigs entre eux en fonction de leur composition nucléotidique et de leur profil d'abondance (Kang et al. 2019; Wu, Simmons, and Singer 2016). Chaque « *bin* » représente alors un ensemble de contigs, *i.*e. un assemblage d'un métagénome (« *Metagenome-Assembled Genome* », MAG).

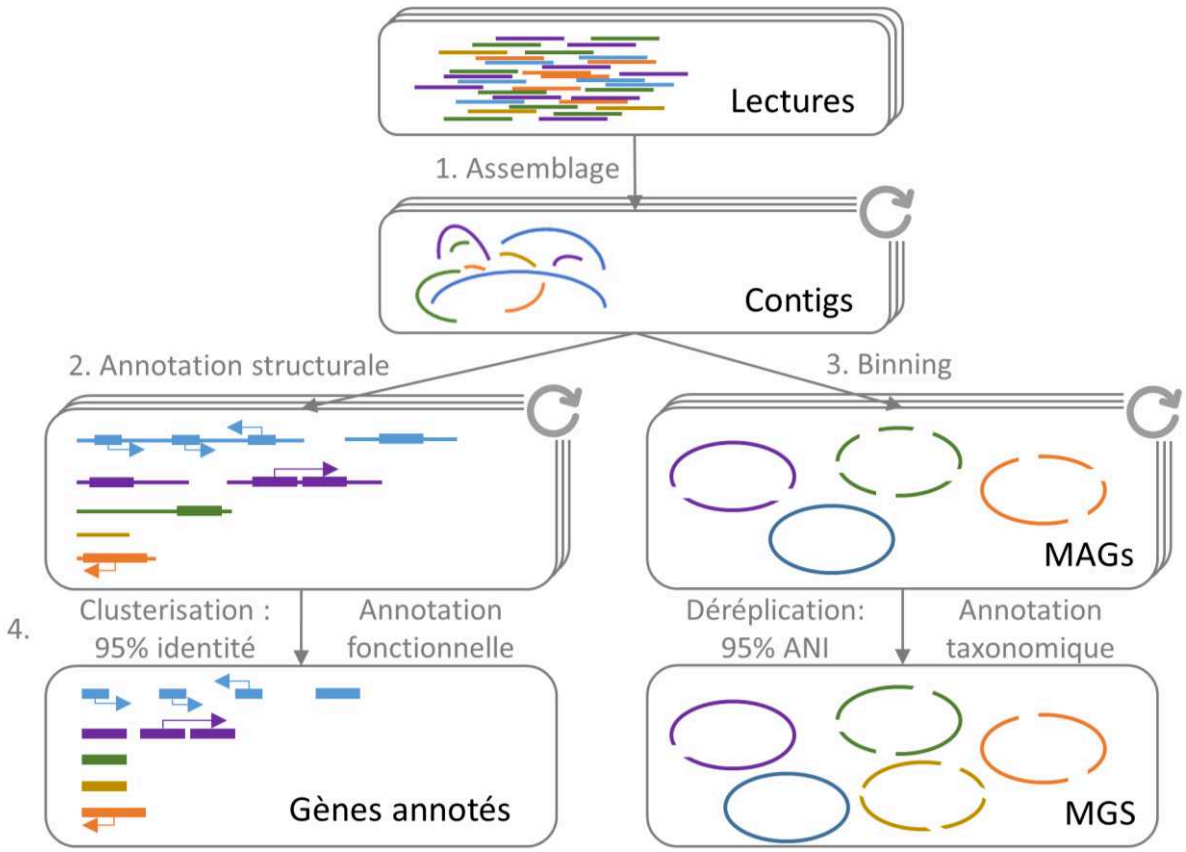


Figure A-12 : Schéma de l'analyse *de novo* des séquences de métagénomique.
Etapes : 1. l'assemblage, 2. l'annotation structurale, 3. le « binning », 4. la réduction de la redondance entre échantillons, et l'annotation fonctionnelle des gènes et taxonomique des MGS. Les flèches circulaires indiquent les étapes qui doivent être faites par échantillon.

Pour définir un catalogue de référence, cette analyse métagénomique doit être faite sur un grand nombre d'échantillons, ce qui permet l'assurance de couvrir au mieux la diversité des génomes présents dans un écosystème. Les dernières étapes du processus d'analyse sont donc de supprimer la redondance entre les MAGs et les gènes identifiés au sein de chaque échantillon. La stratégie est alors de les regrouper selon leur similarité de séquences. Entre gènes, la similarité se calcule par le pourcentage d'identité sur l'ensemble de la longueur du gène (ou au moins sur 90% de leur longueur). Entre MAGs, étant donné le caractère fragmenté et non ordonné de leur séquence (simple ensemble de contigs), la similarité est calculée *via* le pourcentage d'identité moyen (« *Average Nucleotide Identity* », ANI) (Konstantinidis and Tiedje 2005). Le seuil de similarité communément admis est de 95%. En effet, il a été montré que la correspondance entre le seuil de 70% d'hybridation ADN/ADN pour délimiter les espèces bactériennes correspondait à un seuil ANI entre séquence de génome complet oscillant entre 93 et 96% (Rosselló-Móra and Amann 2015). Dans la mesure où, en métagénomique, les assemblages reconstitués sont souvent très fragmentés, le regroupement des MAGs en espèces métagénomiques non redondantes (« *Metagenomic Species* », MGS) ne doit se faire que sur des génomes suffisamment complets et de suffisamment bonne qualité. Pour cela, les MAGs sont évalués grâce à un ensemble de gènes connus pour être universels et présents en une seule copie. La complétude des génomes est évaluée selon le pourcentage de ces gènes retrouvés, et leur contamination en fonction du pourcentage de ces gènes présents en plus d'une copie. Pour être considéré au moins de qualité moyenne, un MAG doit avoir un minimum de 50% de complétude pour un maximum de 10% de contamination, et il est considéré de haute qualité s'il contient plus de 90% des gènes et moins de 5% de ces gènes en multi-copie (Chklovski et al. 2022). Ce seuil de 95% d'identité est également utilisé pour « *clusteriser* » les gènes considérés de même espèce. Il passe à 99% pour la définition de génomes de souches différentes. L'assignation taxonomique se fait généralement à l'échelle des MGS, bien qu'elle puisse être faite à l'échelle des lectures, des contigs ou des gènes. Pour cela, les génomes sont comparés à une base de données de référence, la plus utilisée étant GTDB (« *Genome Taxonomy Database* ») (Parks et al. 2022). L'outil associé à cette base de données, GTDB-tk (Chaumeil et al. 2020), utilise deux ensembles de gènes marqueurs pour déterminer le règne auquel appartient le génome (Bacteria ou Archaea). Il place ensuite le génome dans l'arbre phylogénétique correspondant et attribue une taxonomie en tenant compte de la proximité du génome avec les autres génomes de l'arbre en utilisant un calcul de distance phylogénétique et l'ANI.

Tableau A-1 : Liste des catalogues de microbiotes intestinaux de poule.

Référence	Nombre d'échantillons et types	Conditions	Nombre de gènes / MGS
(Huang et al. 2018)	495 : 5 compartiments intestinaux	7 souches (de chair et de ponte) provenant de 7 fermes chinoises.	9 M de gènes
(Gilroy et al. 2021)	50 + 582 publiés : 5 compartiments intestinaux & fèces	Diverses souches de chair et de ponte, provenant de 12 pays occidentaux, d'Afrique et d'Asie .	20 M de gènes & 5 595 MGS
GG-IGC (Feng et al. 2021)	799 publiés : 5 compartiments intestinaux & fèces	Diverses souches de chair et de ponte provenant de 10 pays (Chine et Europe).	16.6 M de gènes & 1 978 MGS
Projet MetaChick (Plaza Oñate et al. 2023)	340 : cæcum	Diverses souches de chair et de ponte provenant de 30 fermes françaises et 5 systèmes d'élevage.	13.6M de gènes & 2 629 MGS

GG-IGC : « Gallus Gallus Integrated Gene Catalog »; les cinq compartiments intestinaux réfèrent au duodénum, jéjunum, iléon, cæcum et colon.

Le premier catalogue de microbiote intestinal humain a été publié en 2010 (Qin et al. 2010), et pour les animaux d'élevage on retrouve un catalogue pour le microbiote intestinal de porc en 2016 (Xiao et al. 2016), de bovin et de poule en 2018 (Stewart et al. 2018; Huang et al. 2018). Concernant les catalogues liés à la poule, de multiples autres études ont suivi, complétant ou créant ainsi de nouvelles références (**Tableau A-1**). En comparaison aux autres catalogues, celui du projet INRAE MetaChick présente une bonne représentation fonctionnelle des gènes avec un taux d'alignement entre 74.3% et 81.2% des lectures d'échantillons de contenu cæcal séquencés sur les autres projets et tout en représentant une grande diversité d'environnements d'élevages français.

b. Analyse métagénomique sur référence : principe de la métagénomique quantitative

La quantification des micro-organismes et/ou des fonctions correspond à ce que l'on appelle la métagénomique quantitative.

La quantification fonctionnelle se fait par réalignement des lectures sur une référence de gènes. Cette référence peut être un catalogue spécifique de l'écosystème étudié, ou bien une référence plus généraliste (par exemple, les bases de données UniProt et UniRef (Beghini et al. 2021)). L'alignement

est en général contraint par des critères de couverture et d'identité (95% d'identité sur 90% de couverture). L'abondance d'un gène est calculée grâce au nombre de lectures alignées sur celui-ci.

Pour la quantification taxonomique, il existe plusieurs méthodes. La première reprend celle de la quantification fonctionnelle, en réalignant les lectures, cette fois-ci, sur les différents métagénomes (MGS). L'abondance correspond au nombre de lectures, généralement normalisé par la taille des lectures et des génomes pour calculer un nombre de copies de chaque MGS (autrement dit pour les organismes unicellulaires un nombre de cellules). La seconde méthode se base quant à elle sur le principe de gènes marqueurs. L'abondance d'un génome sera calculée par l'abondance moyenne ou médiane de ces gènes. La définition des gènes marqueurs peut être pré-calculée (comme dans l'outil MetaPhlan4) (Blanco-Miguez et al. 2022) ou bien être calculée *de novo* (avec les outils Canopy ou MSPminer (Nielsen et al. 2014; Plaza Oñate et al. 2019)). Pour cette analyse *de novo*, les gènes sont alignés sur les génomes, et ils sont définis comme gène marqueur, en fonction de leur prévalence dans les MAGs composant une MGS et en fonction de leur co-abondance dans les différents échantillons étudiés. Cette stratégie a été développée initialement comme une méthode alternative au « *binning* ». Les génomes ne sont alors pas reconstitués *via* un regroupement des contigs mais par regroupement des gènes co-abondants. On parle alors de construction de pan-génome d'espèce métagénomique (« *Metagenomic Species Pan-genome* », MSP).

c. Biais et difficultés liés à l'analyse de séquences métagénomiques

Comme mentionné précédemment, le frein principal à l'utilisation de la métagénomique est son coût financier, ainsi que la difficulté d'analyse des données. En effet, l'avantage de cette technique sur le séquençage métabarcoding est une augmentation de la sensibilité et de la résolution taxonomique et fonctionnelle des résultats (Bharti and Grimm 2021). Mais pour obtenir ces résultats, le séquençage doit être fait à une grande profondeur. Toujours d'un point de vue du protocole expérimental, la métagénomique peut souffrir d'une forte contamination selon le type d'écosystème étudié. Par exemple, le séquençage d'écosystèmes où le prélèvement d'une biomasse suffisante est difficile, peut être contaminé par de potentiels micro-organismes présents dans les réactifs (Quince et al. 2017; Bharti and Grimm 2021). D'autres écosystèmes sont eux naturellement « contaminés » par des cellules de l'hôte. Cette contamination peut même représenter la majorité des cellules, par exemple plus de 99% dans le microbiote nasal ou vaginal, alors qu'elle représente moins de 1% dans les fèces humains (Gevers et al. 2012). Pour réduire ces contaminants, plusieurs protocoles sont disponibles permettant notamment l'enrichissement du signal provenant des micro-organismes (Quince et al. 2017). Par ailleurs, bien qu'une chaîne de traitements de ces séquences commence à se standardiser, le choix des

outils à utiliser, dans quel contexte et avec quelle référence est encore loin d'être évident. De plus, traiter une si grande quantité de données nécessite des ressources humaines et informatiques (calculs et stockage) importantes. Avec l'accumulation des données de séquences stockées dans les bases de données qui permettent d'accroitre la quantité et la qualité des connaissances, il est possible de limiter la profondeur de séquençage pour une analyse s'appuyant uniquement sur les génomes, gènes et annotations d'écosystèmes déjà bien caractérisés. Toutefois pour des écosystèmes complexes comme celui du microbiote intestinal, même largement étudié comme celui de l'homme, de nouveaux séquençages ne cessent de faire grossir les catalogues de référence avec de nouveaux gènes et génomes. Ainsi, Kim *et al.* ont montré que bien que composé de près de 205 000 MGS, le catalogue de référence des génomes du microbiote intestinal humain (Almeida et al. 2021) sous-représentait la diversité microbienne présente dans certaines populations asiatiques de Corée, du Japon et d'Inde (Kim et al. 2021). L'analyse *de novo*, impliquant une profondeur de séquençage importante, est donc souvent priorisée et ceci est d'autant plus vrai pour l'analyse d'écosystèmes moins bien caractérisés comme celui de la poule. Cette stratégie permet d'être plus exhaustif sur la caractérisation des gènes, fonctions et micro-organismes présents.

Que ce soit le séquençage d'un gène marqueur (métabarcoding), ou de l'intégralité de l'ADN (métagénomique), leurs analyses ne peuvent que révéler la présence des micro-organismes et éventuellement leur fonction dans les milieux étudiés, mais ne peuvent pas présager de leur activité. Pour mesurer l'activité réelle de ces communautés microbiennes il est nécessaire de compléter ces analyses avec, par exemple, le séquençage des ARN *i.e*. la métatranscriptomique.

A.5.3. La métatranscriptomique : séquençage complet des transcriptomes

Alors que les analyses précédentes, métabarcoding et métagénomique, permettent de caractériser avec plus ou moins de précision les communautés microbiennes d'un écosystème et les fonctions portées par ces organismes, la métatranscriptomique permet de mettre en évidence, à un instant T, les fonctions actives, c'est-à-dire celles issues des gènes exprimés. De plus, les analyses sur l'ADN, en particulier la métagénomique, souffrent de la présence d'ADN reliques, des molécules d'ADN extra-cellulaires provenant de la dégradation de cellules mortes (Berg et al. 2020). Ces molécules peuvent représenter une proportion importante de l'échantillon séquencé, par exemple, 44% dans des échantillons de sol, et en moyenne 33% et jusqu'à 80% dans divers écosystèmes environnementaux ou intestinaux (Carini et al. 2016; Lennon et al. 2018). Par ailleurs, la présence d'un gène ne présage pas de son activité au cours du temps ou des conditions. La quantification fonctionnelle obtenue sur l'ADN ne représente donc que le potentiel de ce que peuvent faire les micro-organismes, alors que la

métatranscriptomique mesure ce que les micro-organismes font, tout du moins à l'échelle de l'expression des gènes.

Le processus d'analyses de l'expression des gènes d'un microbiote suit les grandes étapes du protocole d'analyse d'un transcriptome unique (analyse RNASeq). L'ARN total est extrait, les ARN messagers (ARNm) sont ciblés, puis séquencés. La quantification de l'expression des gènes peut alors se réaliser de deux manières soit *de novo*, soit vis-à-vis d'une référence connue. Mais, comme pour la métagénomique, bien que ces principes soient globalement bien maîtrisés lorsque l'on analyse un unique transcriptome, la complexité d'un écosystème microbien ajoute des difficultés importantes à la fois au niveau du protocole expérimental et au niveau du traitement des données.

a. Protocole expérimentale et d'analyse : focus sur la sélection des ARN messagers.

Tout comme dans le cadre d'une analyse de transcriptome d'un seul organisme, le protocole expérimental nécessite d'amplifier le signal des ARNm par rapport à celui des ARNr (**Figure A-13**). En effet, bien que les proportions des différents types de transcrits fluctuent selon l'activité cellulaire, les ARN ribosomaux représentent la très grande majorité des transcrits (>80%), suivis des ARN de transfert (~10%), des ARN messagers (~5%), et l'ensemble des autres ARN non-codants constitue le reste (Westermann and Vogel 2021). Si les ARNr étaient conservés, un séquençage de trop grande profondeur serait donc nécessaire pour analyser les autres transcrits. Dans le cadre de l'analyse de transcriptome d'animaux d'élevage (donc eucaryotes), la sélection des ARNm se fait généralement par enrichissement. Cet enrichissement se fait grâce à l'utilisation de sondes oligo-dT (courtes séquences composées exclusivement de thymines) qui vont cibler l'extrémité 3' poly-adénilée (poly-A) des ARN messagers. Dans un contexte d'analyse d'un microbiote composé essentiellement de bactéries, comme les microbiotes intestinaux, la technique d'enrichissement par oligo-dT ne peut être utilisée, les transcrits procaryotes ne possédant que rarement de queue poly-A. Il est alors nécessaire d'utiliser des stratégies de suppression ou de blocage des ARNr, la ribodéplétion. Un mélange d'oligonucléotides est défini pour cibler les ARNr qui sont ensuite supprimés par digestion enzymatique, capturés par des billes magnétiques ou bloqués pour ne pas être reverse-transcrits (Wahl, Huptas, and Neuhaus 2022). C'est une étape délicate car l'ARN est une molécule plus fragile que l'ADN, l'extraction doit donc aboutir à des ARN de haute qualité pour optimiser l'efficacité de la ribodéplétion. Elle est par ailleurs complexifiée dans le cadre d'expérience de métatranscriptomique à cause de la multiplicité des organismes qui doivent être ciblés. De plus ,tout comme en métagénomique, les ARNm peuvent être « contaminés » par les ARNm de l'hôte. Le contrôle de cette contamination peut se faire

expérimentalement grâce à l'hybridation sélective utilisant par exemple des oligo-dT ciblant les ARNm de l'hôte eucaryote et permettant l'enrichissement de transcrits non poly-A procaryotes.

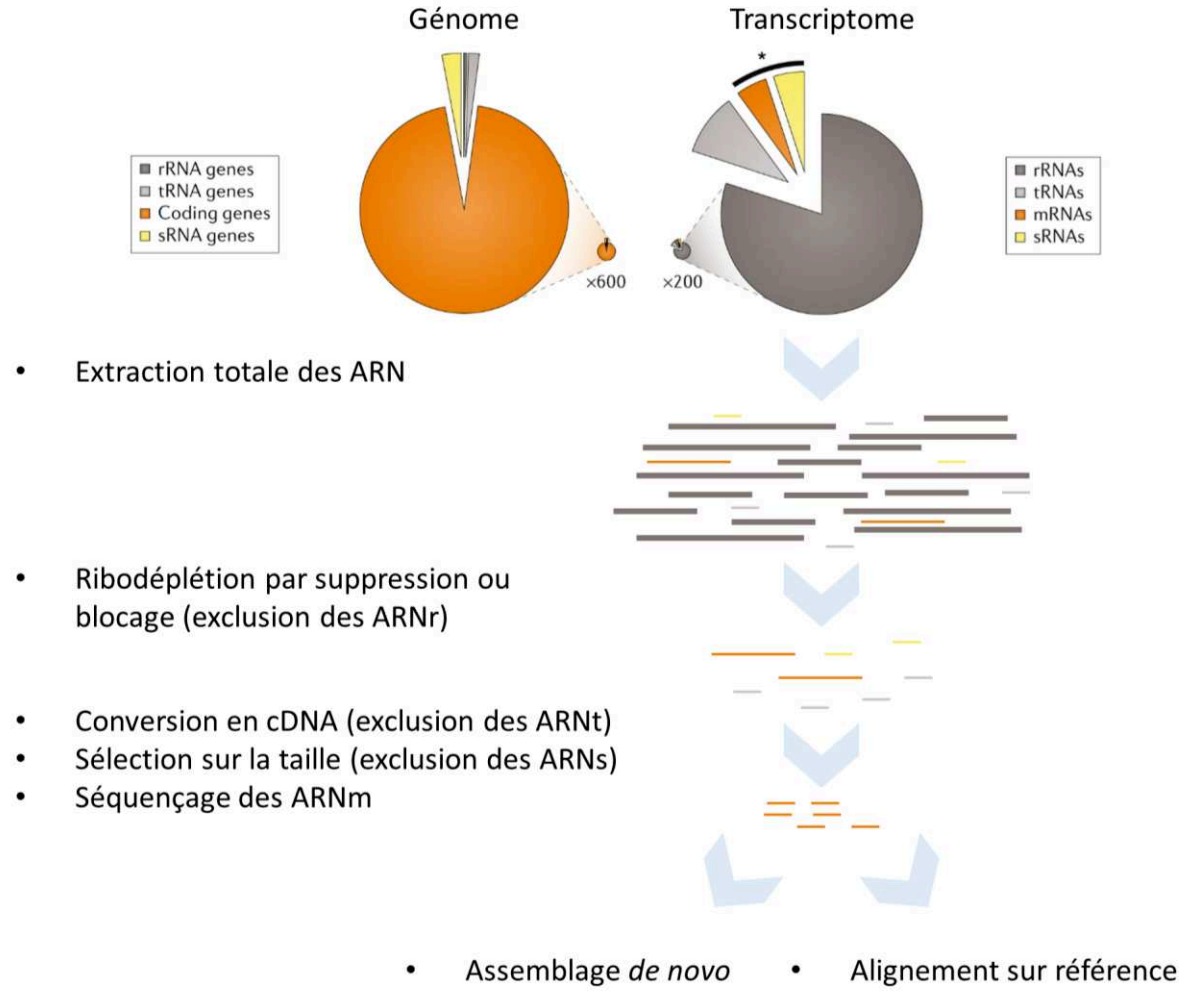


Figure A-13 : Schéma du protocole expérimental de métatranscriptomique.
Les diagrammes circulaires représentent la fraction des différentes classes de gènes dans le génome (à gauche) ou des molécules d'ARN dans le transcriptome chez une souche de *Salmonella enterica* (figure adaptée de (Westermann and Vogel 2021). rRNA = ARNr : ARN ribosomaux ; tRNA = ARNt : ARN de transfert ; sRNA = ARNs : petits ARN non-codants ; mRNA = ARNm : ARN messagers.

L'analyse bioinformatique des séquences suit les mêmes principes qu'en métagénomique. Les lectures doivent être nettoyées, notamment pour contrôler les contaminations restantes des séquences par les ARNr ou provenant de l'hôte. Elles peuvent ensuite être assemblées pour former une référence de transcrits, analyse *de novo,* ou bien être directement comparées à une référence de gènes connus, analyse quantitative. L'assemblage de transcrits présente des défis supplémentaires à l'assemblage de génome, notamment à cause de leur variabilité d'expression qui rend leur représentativité au sein des lectures séquencées encore moins homogènes. Cette représentativité déséquilibrée est exacerbée en

métatranscriptomique puisqu'elle est également influencée par la variabilité d'abondances des micro-organismes actifs. L'analyse s'appuyant sur une référence connue peut se baser sur le résultat d'une analyse de métagénomique correspondant au même écosystème, ou bien sur les bases de référence généralistes en utilisant les mêmes outils qu'en métagénomique (principalement de l'alignement de séquences).

b. Biais et difficultés liés à l'analyse de métatranscriptomique

La métatranscriptomique est une méthode encore récente pour laquelle nous avons peu de retour d'expérience. La standardisation des protocoles, des outils et des stratégies d'analyses sont ainsi encore en développement. Les premières difficultés de cette méthode relèvent des propriétés intrinsèques de la molécule d'ARN, le contexte de l'analyse d'un microbiote au lieu d'un organisme unique exacerbant ces difficultés. Les kits et procédures de chacune des étapes aboutissant au séquençage du métatranscriptome peuvent influencer la composition observée *in fine*. En effet, l'ARN est une molécule fragile qui se dégrade plus rapidement que la molécule d'ADN. Après le prélèvement, il est recommandé de la protéger contre les enzymes qui la dégradent et selon la méthode de conservation des échantillons la dégradation est plus ou moins bien stoppée (Westermann and Vogel 2021). Par ailleurs, et comme indiqué précédemment, les ARN messagers ne représentent qu'une faible proportion des ARN totaux et peuvent en plus être contaminés par ceux des cellules de l'hôte. Les protocoles de ribodéplétion et de déplétion des ARNm de l'hôte doivent être également évalués selon les écosystèmes. Ces étapes supplémentaires, nécessitant parfois des kits dédiés à l'écosystème étudié, compliquent la reproductibilité des analyses et engendrent des coûts financiers d'expérimentation supplémentaires.

Au niveau de l'analyse, en particulier d'écosystèmes jusque-là mal caractérisés, l'assemblage du métatranscriptome est indispensable pour permettre la plus grande exhaustivité des transcrits, mais cette analyse *de novo* n'est pas une étape facile, et il est souvent recommandé de travailler sur une référence connue ou en combinaison avec un protocole de séquençage de métagénomique. Enfin, même si cette méthode permet indéniablement d'avoir une vue fonctionnelle plus réaliste de l'activité d'un écosystème microbien que celle obtenue sur les données de métagénomique, elle ne représente pas l'activité métabolique et protéique réelle (Bashiardes, Zilberman-Schapira, and Elinav 2016). Pour avoir une vue, complète de l'activité du microbiote, des analyses de métaprotéomique et métabolomique sont à envisager.

B. OBJECTIFS DE LA THESE

Nous avons vu dans l'introduction précédente que la poule et notamment la poule pondeuse représentait une espèce majeure parmi les animaux d'élevage. Les œufs représentent la source alimentaire d'origine animale la moins onéreuse au monde ce qui explique sa forte augmentation de production, en particulier dans les pays en voie de développement, et son rôle dans le maintien de la sécurité alimentaire dans les régions les plus pauvres. Pour répondre à cette demande croissante, de nombreuses études ont été réalisées pour améliorer l'efficience alimentaire de ces animaux. Ces recherches ont permis aux entreprises de sélection de produire des animaux désormais efficients et de les distribuer partout dans le monde. Pour autant, cette efficience est particulièrement dépendante de l'environnement dans lequel les animaux de production sont élevés. Dans un contexte de changement climatique fort et de conditions géopolitiques instables, l'amélioration de l'efficience alimentaire reste un objectif scientifique, économique et écologique majeur. L'identification des leviers qui l'influencent permettrait de répondre aux nouveaux enjeux de limitation de l'impact des élevages sur l'environnement, de réduire la compétition entre la nourriture à destination des animaux et celle à destination des hommes, et d'améliorer l'adaptation des animaux à une modification de la composition du régime alimentaire.

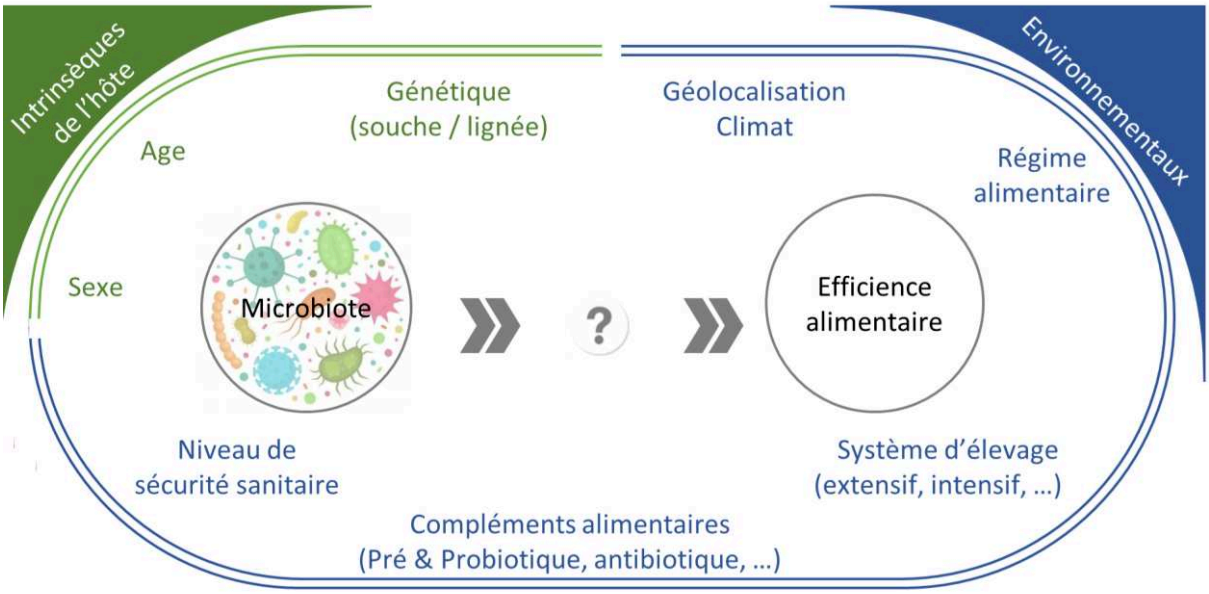


Figure B-1: Facteurs d'influence du microbiote intestinal et de l'efficience alimentaire des poules.

Parmi ces leviers, le microbiote intestinal, et cæcal en particulier, représente une composante de plus en plus étudiée. Son rôle primordial dans la dégradation des aliments, l'absorption des nutriments, et

la production de métabolites influençant le métabolisme de l'hôte alimente largement la recherche de cette dernière décennie. Cependant, cet écosystème microbien est connu pour être sensible à une multitude de facteurs intrinsèques de l'hôte et environnementaux, dont un grand nombre influencent également l'efficience alimentaire (**Figure B-1**). Il est donc nécessaire de multiplier les analyses dans des contextes variés pour mieux comprendre le lien potentiel entre microbiote et efficience alimentaire.

Par ailleurs, des progrès méthodologiques importants ont été réalisés ou sont en cours dans l'analyse des communautés microbiennes et de leur interaction avec leur hôte. Chaque méthodologie présente des avantages et inconvénients techniques mais répond également de façon complémentaire aux questions de recherche.

C'est dans ce contexte à la fois scientifique et technique que se positionne cette thèse. Elle a vocation à répondre à trois grandes questions :

Le microbiote a-t-il un rôle sur l'efficience alimentaire de la poule pondeuse ?

Le microbiote intestinal de poule pondeuse est largement sous-étudié en comparaison aux microbiotes intestinaux d'autres animaux d'élevage, poulet de chair inclus. Bien que plusieurs études laissent penser que le microbiote intestinal pourrait influencer l'efficience alimentaire, l'association entre la composition du microbiote et l'efficience alimentaire n'est à ce jour toujours pas établie. De plus, le microbiote intestinal est souvent caractérisé par l'identification des communautés microbiennes qui le composent, mais peu d'études vont jusqu'à l'identification de leurs fonctions. Ainsi, est ce que l'hypothèse d'une association entre le microbiote et l'efficience peut également s'observer à l'échelle des fonctions portées par le microbiote ?

Quelle est l'influence du régime alimentaire sur le microbiote et sur son potentiel rôle dans l'efficience alimentaire ?

Nous avons choisi d'évaluer l'influence du régime alimentaire sur le microbiote intestinal. Puisque le rôle du microbiote est notamment de dégrader les aliments, nous nous attendons à ce que la composition microbienne varie selon les modifications du régime alimentaire. Au-delà de la composition microbienne, est ce que les fonctions associées sont également impactées ? Enfin, est ce que le lien potentiel entre efficience alimentaire et microbiote est maintenu quel que soit le régime alimentaire ?

Quel type de données omiques pour répondre à quelle question ?

En sus des objectifs scientifiques mentionnés précédemment, nous nous sommes intéressées lors de cette thèse aux méthodologies permettant l'analyse du microbiote. Le séquençage métabarcoding est largement utilisé mais limite l'analyse du microbiote du fait de sa plus faible résolution taxonomique et par une analyse fonctionnelle possible par inférence uniquement. Les techniques de métagénomique et de métatranscriptomique engendrent de beaucoup plus larges quantités d'informations et devraient permettre d'obtenir une image plus précise des taxonomies et des fonctions. En particulier, la métatranscriptomique permettra de mesurer les fonctions actives et pas seulement potentielles de ce microbiote. Dans cette thèse, nous tenterons de comparer ces trois méthodologies du point de vue des résultats biologiques, ainsi que des contraintes, difficultés et bénéfices que chacune apporte.

C. DEMARCHE EXPERIMENTALE ET ACQUISITION DES DONNEES

C.1. Des lignées divergentes pour l'efficience alimentaire comme modèle expérimental

L'analyse du déterminisme d'un caractère nécessite d'obtenir un ensemble d'individus présentant une variabilité phénotypique pour ce caractère. La comparaison des individus extrêmes d'une même population peut ensuite aider à comprendre les mécanismes physiologiques et génétiques déterminant ce caractère. Pour être efficace, cette comparaison doit s'effectuer sur des groupes d'individus (ici des poules) de phénotypes opposés différant fortement entre eux, ce qui est souvent difficile à obtenir dans une même population. C'est pourquoi l'utilisation de lignées obtenues par sélection divergente représente un outil puissant. En effet, génération après génération, les lignées présentent une différence de plus en plus marquée du caractère sous sélection. Dans notre étude, nous avons utilisé deux lignées de poules pondeuses, R+ et R-, sélectionnées de manière divergente depuis 1976 sur leur efficience alimentaire (mesurée par la RFI) à partir d'une même population de la souche Rhode Island Red (6 coqs et 50 poules). A chaque génération et pour chaque lignée, 9 coqs R- et 10 R+ ont été croisés avec 45 ou 50 poules (5 poules par coq) pour obtenir une population en sélection constituée de 40 coqs et 160 poules adultes (Bordas and Mérat 1984; Bordas, Tixier-Boichard, and Merat 1992).

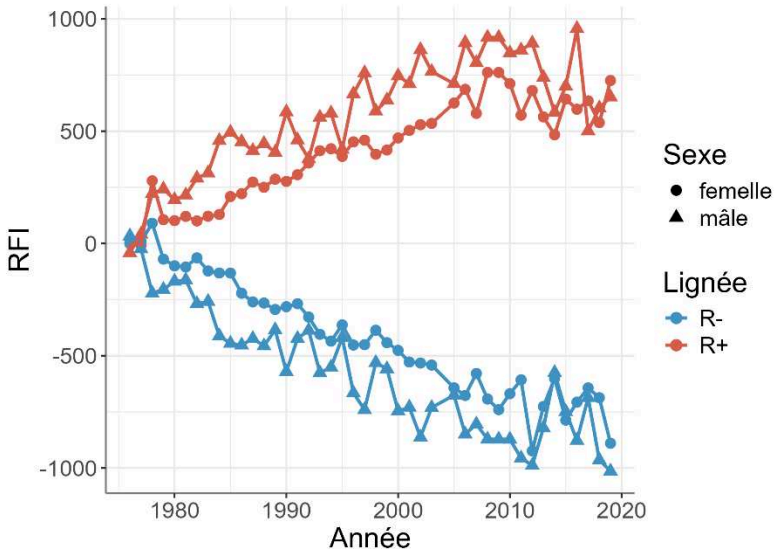


Figure C-1: Evolution de la consommation résiduelle d'aliments (RFI) des lignées R+ et R-.

Comme indiqué en introduction (paragraphe **A.2.1**), la RFI correspond à la différence entre la quantité d'aliment ingéré observée et celle attendue. Ainsi, la lignée R-, avec une RFI négative est dite efficiente,

et la lignée R+, avec une RFI positive est dite non efficiente. Après presque 50 ans de sélection phénotypique, les deux lignées diffèrent de plus de 5 écarts-types phénotypiques sur leur mesure de RFI (**Figure C-1**).

Parallèlement à cette évolution directe attendue du RFI, plusieurs autres caractères ont également évolué de manière divergente, alors que le poids ou la production des œufs restent similaires entre les deux lignées. On peut citer des caractères directement liés à l'efficience alimentaire. Ainsi, la prise alimentaire est plus de 1,5 fois pour les poules et 2 fois pour les coqs plus importante chez la lignée non efficiente R+ que chez la lignée efficiente R-. L'indice de consommation (FCR), avec une prise alimentaire plus importante pour une production similaire d'œufs, est logiquement plus importante pour la lignée R+ que pour la lignée R-. On peut également citer des caractères indirectement sélectionnés qui caractérisent un métabolisme énergétique différent entre les deux lignées. La lignée R+ a ainsi des tarses ("tige" des pattes), une crête et des barbillons plus grands permettant une meilleure évacuation de la chaleur corporelle qui est plus importante du fait d'une thermogénèse induite par l'alimentation plus importante (Gabarrou et al. 1997; 1998; Tixier-Boichard et al. 1995). Malgré une prise alimentaire réduite chez la lignée R-, la quantité de tissus adipeux (qui permet un stockage de l'énergie) est, quant à elle, plus importante que chez la lignée R+, avec notamment chez les femelles une quantité de lipide dans le foie 2,7 fois plus élevée (El-Kazzi et al. 1995b). Plus récemment, il a été montré que la lignée R+ présente une meilleure résistance à la bactérie *Escherichia coli* (responsable de colibacilose), ainsi qu'une meilleure réponse vaccinale (Quéré 2017; Zerjal et al. 2021). Ceci suggère des compromis différents d'utilisation de l'énergie entre fonctions métaboliques, dans chacune des deux lignées (Zerjal et al. 2021).

C.2. Mise en place de l'expérimentation animale et échantillonnage

Les deux lignées divergentes décrites précédemment ont été utilisées au sein du projet ANR ChickStress (2014-2019) qui visait à mieux comprendre la base génétique de l'adaptation à la chaleur et à un régime alimentaire de moins bonne qualité nutritive. Dans un contexte de changement climatique, de mondialisation de la production de souches commerciales, et de compétition entre alimentation animale et humaine pour certaines ressources céréalières, mieux comprendre les mécanismes impliqués dans les capacités adaptatives des animaux d'élevage est un enjeu majeur. L'ensemble du dispositif expérimental du projet ChickStress inclut des poules de quatre lignées. En plus des poules des lignées R+ et R- présentées précédemment, des poules des lignées Fayoumi et LS ont été utilisées. Ces lignées correspondent respectivement à une lignée de poule ancienne originaire

d'Egypte et une lignée expérimentale INRAE naine et porteuse de la mutation « cou nu » qui détermine l'absence de plumes au niveau du cou. Elles présentent toutes deux une forte résistance à la chaleur contrairement aux lignées R+ et R-. Lors de ce projet, des poules des quatre lignées sont nées et ont été élevées ensemble jusqu'à l'âge adulte (17 semaines), puis placées en cage individuelle. Elles ont été nourries avec un régime optimal à une température ambiante de 22° (condition référencée CTR pour « Control » dans la suite du manuscrit) ou bien, pendant les 4 dernières semaines d'expérimentation, à une température ambiante de 32° (condition référencée HS pour « *Heat Stress* » dans la partie **C.3.4**). Enfin, des poules des lignées R+ et R- ont été nourries avec un régime appauvri en énergie (condition référencée LE pour « Low-Energy » dans la suite du manuscrit) sans modification de la température ambiante.

Notre étude a bénéficié de ce dispositif en se focalisant sur les données issues de la partie « adaptation à une modification du régime alimentaire » (**Figure C-2**, **Tableau C-1**). Les poules utilisées dans ce dispositif (condition CTR et LE) représentent l'ensemble des familles de pères de chacune des lignées R+ et R-.

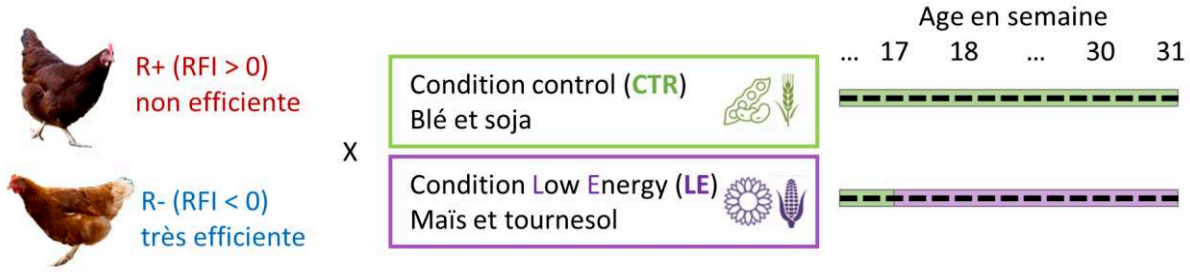


Figure C-2 : Représentation du plan d'expérimentation animale issu du projet ChickStress.

Tableau C-1 : Résumé de la composition en céréales des deux régimes alimentaires.

	CTR	LE
Blé	599	180
Maïs	50	332.5
Soja	197.6	92.1
Tournesol	0	152.9
Colza	20	50
Avoine	0	55.7

CTR correspond au régime contrôle et LE au régime appauvri en énergie. La composition détaillée est décrite dans le Tableau 4 de l'article publié pendant cette thèse (Bernard et al. 2024) et inclus dans le paragraphe **D.5**.

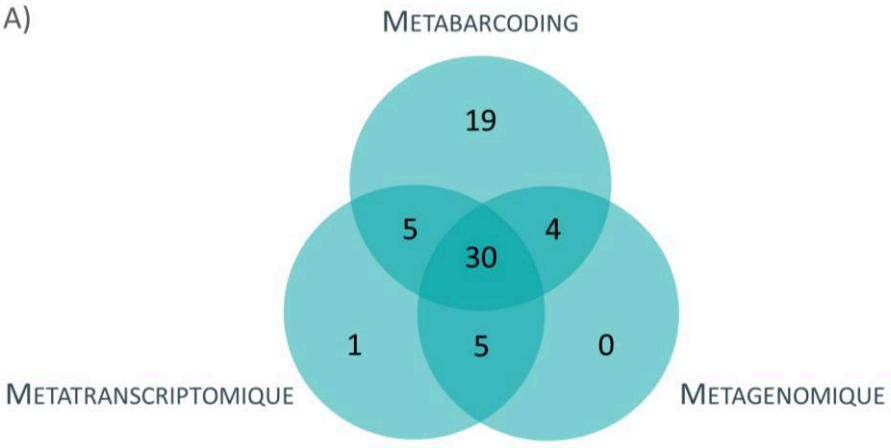
Le régime CTR correspond à un régime commercial optimisé pour les poules pondeuses adultes, c'est-à-dire riche en amidon et contenant 3% de fibres indigestes. La formulation du régime LE se rapproche de celle que l'on pourrait trouver dans les pays d'Asie ou d'Afrique. Il contient 15% d'énergie métabolisable en moins, ce en réduisant les céréales riches en amidon (blé et maïs) et en augmentant les céréales riches en fibres indigestes (tournesol, colza et avoine). Bien que ces deux régimes soient différents du point de vue de leur teneur en énergie, amidon et fibre, ils sont similaires du point de vue du contenu protéique.

A 31 semaines d'âge, les animaux ont été sacrifiés et leur contenu cæcal a été prélevé en vue de la présente étude sur les liens entre microbiote, efficience alimentaire et adaptation à une modification du régime alimentaire.

C.3. Acquisition des données de séquences et processus d'analyses de données

Les contenus cæcaux d'une soixantaine d'animaux ont été utilisés pour trois expériences de séquençage différentes : le métabarcoding (58 poules), la métagénomique complète (39 poules) et la métatranscriptomique (41 poules) (**Figure C-3**).

A)



B)

	R- CTR	R+ CTR	R- LE	R+ LE
Metabarcoding	18	18	11	11
Métagénomique	10	10	10	9
Métatranscriptomique	11	11	9	9

Figure C-3: Nombre d'animaux séquencés par type de données omiques.

Pour rappel, ces animaux ont été produits lors du projet ChickStress qui s'intéressait à la fois à l'impact d'un stress thermique ou d'une modification du régime alimentaire sur différentes lignées de poule pondeuses. Malheureusement un défaut d'éclosion a eu lieu, et il a été décidé lors de ce projet de mettre la priorité sur l'analyse du stress thermique. Les échantillons CTR servant de contrôle pour l'analyse des deux stress, cela explique pourquoi il y a plus d'animaux dans notre condition CTR que notre condition LE.

Chacune de ces méthodes s'est reposée sur la technologie de séquençage Illumina, méthode à haut débit à ce jour la plus fidèle avec un taux d'erreur de l'ordre de 0.08 à 0.6% selon les séquenceurs (Stoler & Nekrutenko, 2021). L'analyse bioinformatique de ces séquences vise à atteindre trois objectifs : 1) reconstruire un ensemble de séquences non redondantes représentant la diversité taxonomique présente dans les cæca, 2) annoter taxonomiquement et fonctionnellement cette diversité, 3) quantifier à la fois les taxonomies et les fonctions selon les conditions expérimentales.

C.3.1. Caractérisation du microbiote cæcal par séquençage ciblé de l'ARNr 16S

a. Acquisition des séquences de métabarcoding

Comme indiqué en introduction, le métabarcoding ou séquençage ciblé d'un gène marqueur consiste à séquencer une région de l'ADN qui doit être universelle, informative, amplifiable et répondre aux contraintes de taille de fragment imposées par la technologie de séquençage utilisée.

Pour le choix de ce gène marqueur, dans la mesure où le microbiote cæcal de poule est majoritairement composé de bactéries, nous avons ciblé le gène de l'ARN ribosomal 16S (ARNr 16S). Ce gène, présent dans l'ensemble du règne des procaryotes représente le marqueur le plus communément utilisé pour les analyses de microbiotes intestinaux notamment chez les animaux d'élevage (Deusch et al. 2015).

Le choix de la région du gène de l'ARNr 16S ne fait pas consensus car leur capacité à discriminer les espèces microbiennes est variable selon les rangs taxonomiques des bactéries ciblées (Bindari and Gerber 2022; Johnson et al. 2019) (**Figure C-4**). La région V4 par exemple, est la plus longue avec une faible variabilité de taille. Elle a été une des plus utilisées dans les premières études de métabarcoding. On sait aujourd'hui que son pouvoir discriminant est en fait faible en comparaison aux autres régions de l'ARNr 16S (Johnson et al. 2019). Avec l'augmentation des tailles de lectures séquencées, l'utilisation de régions plus longues, avec donc un potentiel discriminant plus important, a été possible, mais chaque région a ses spécificités qui par ailleurs dépendent de l'écosystème étudié. La région V3-V4 permet de détecter une plus grande diversité bactérienne que la région V4-V5 (Rintala et al. 2017),

et aboutit à des résultats similaires à la région V3-V5 (Darwish et al. 2021) pour des analyses d'écosystèmes intestinaux. Les amorces utilisées dans notre étude ciblent cette région (Nadkarni et al. 2002), et les amplicons générés, d'une taille allant de 430 à 460 pb, ont été séquencés grâce une librairie pairée de 2x250 pb sur un séquenceur Illumina MiSeq.

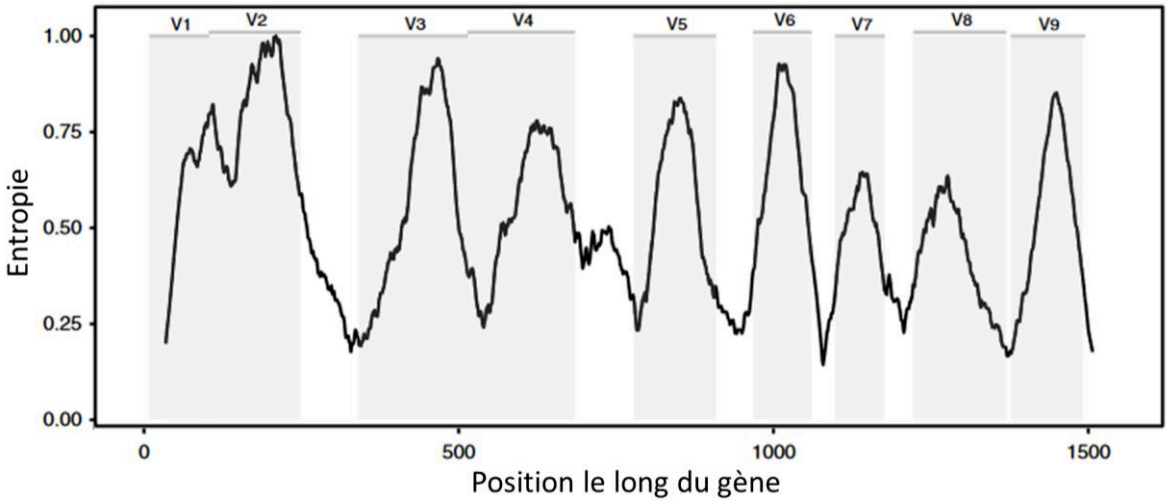


Figure C-4 : Entropie et taille des régions hypervariables du gène de l'ARNr 16S.
L'entropie représente la diversité nucléotidique. Figure adaptée Johnson et al. 2019.

b. Méthode et outils d'analyse bioinformatique des séquences de métabarcoding

L'analyse de ces séquences a été réalisée avec la suite d'outils FROGS (Escudié et al. 2018; Bernard et al. 2021) développée depuis 2014 à INRAE et dont je suis la principale développeuse. La suite repose sur un processus classique d'analyse de séquences d'amplicons : nettoyage des lectures et assemblage des amplicons dans le cas de lectures pairées, *clustering* des séquences, détection et suppression des chimères, filtres des clusters générés, annotation taxonomique et inférence fonctionnelle (**Figure C-5**).

Lors de l'étape de nettoyage, les lectures sont assemblées avec PEAR (J. Zhang et al. 2014), les amorces de PCR sont enlevées et les amplicons sont filtrés sur leur composition nucléotidique (absence de N) et sur leur taille. Les paires de lectures sont assemblées et sont ensuite clusterisés par l'outil swarm (Mahé et al. 2014). Contrairement aux outils de *clustering* classiques, swarm utilise un nombre maximum fixe de différences entre deux séquences (ici d=1) pour les regrouper en cluster. Chaque cluster aura donc une similarité de séquence globale variable et non un figée à 97% ou 99% d'identité. Les séquences représentatives des clusters sont ensuite filtrées selon leur statut chimérique et sur leur abondance relative globale qui doit être > 5.10^5 pour éviter la prise en compte de séquences artéfactuelles (Bokulich et al. 2013). Une des particularités de FROGS est de faire une validation croisée de la détection des chimères limitant ainsi le taux de faux négatifs. Les séquences concernées doivent être détectées comme chimériques dans l'ensemble des échantillons dans lesquels elles sont

présentes pour être effectivement considérées comme chimériques. Dû à l'utilisation de swarm qui permet une plus grande discrimination des séquences, et à l'application des filtres, FROGS produit des séquences non redondantes se rapprochant des ASV.

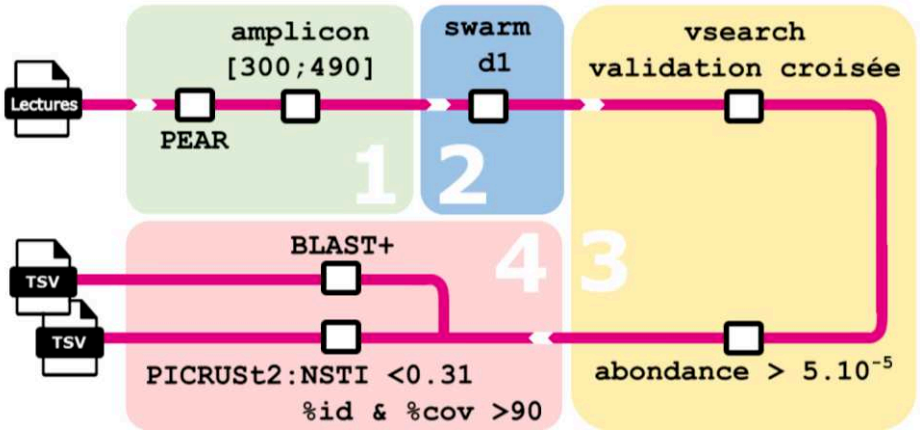


Figure C-5: Représentation graphique des outils et paramètres de FROGS utilisés dans cette étude. Les briques correspondent à : 1. nettoyage des lectures et assemblage des amplicons, 2. *clustering*, 3. filtres des chimères et des amplicons rares, 4. annotation taxonomique et inférence fonctionnelle des ASV.

L'annotation taxonomique repose sur la disponibilité d'une base de référence de qualité et exhaustive. Pour notre étude, nous avons fait le choix d'utiliser l'outils d'alignement de séquence BLAST+ sur la base de référence SILVA (Quast et al. 2013). Cette base est une des bases de séquences du gène de l'ARNr 16S les plus complètes. Elle est régulièrement mise à jour et largement utilisée dans les analyses d'écosystèmes microbiens. Elle propose par ailleurs un indice de confiance, le *pintail*, allant de 0 à 100. En utilisant un seuil minimal de 50, cela permet de limiter les annotations dues à des séquences douteuses du point de vue de leur qualité intrinsèque mais également de leur ressemblance phylogénétique avec les autres séquences de la base. La version 138.1, filtrée sur un score de *pintail* de 50 et utilisée lors de notre étude, contient 385 381 séquences du gène de l'ARNr 16S correspondant à 3 613 genres de bactéries et d'archées. Depuis la version 4.0.0 de FROGS, sortie en 2022 et dont j'ai supervisé le développement pendant cette thèse (Darbot et al. 2022, également disponible en annexe **J.1**), FROGS inclut une annotation fonctionnelle des ASV par inférence en reposant sur la suite d'outils PICRUSt2 (Douglas et al. 2020). Cette suite permet d'inférer les fonctions présentes, en attribuant à chaque ASV, les fonctions de l'organisme le plus proche phylogénétiquement. La base de connaissances de PICRUSt2 (un arbre phylogénétique du gène de l'ARNr 16S), mise à jour en 2018 à partir de 41 926 génomes complets de la base de données IMG (« *Integrated Microbial Genome database* »), inclut 20 000 séquences d'ARNr 16S pleine longueur, correspondant à 1 925 genres de bactéries et d'archées.

Le chapitre **D** expose les analyses de ces séquences de métabarcoding obtenues sur les prélèvements cæcaux des deux lignées de poules nourries avec les deux régimes. Cette étude donne une première image des différences de composition microbienne et de fonction associées à l'efficience alimentaire et/ou à chacun des deux régimes.

C.3.2. Caractérisation du microbiote cæcal par séquençage complet de l'ADN

a. Acquisition des séquences de métagénomique

Le séquençage complet du microbiote cæcal fait partie intégrante du projet France Génomique, MetaChick (2015), soutenu par différentes entreprises avicoles réunies en consortium avec plusieurs unités INRAE et l'ITAVI, en collaboration avec la plateforme de séquençage du Génoscope (CEA, Evry). Ce projet s'est déroulé en deux phases répondant à deux objectifs différents.

La phase 1 du projet qui a démarré avant cette thèse, a consisté à séquencer les microbiotes de contenus cæcaux d'une grande diversité d'échantillons provenant de poulets de chair (mâle et femelle) à différents âges, et de poules pondeuses adultes, de différentes souches et lignées, et élevés dans des fermes commerciales ou expérimentales en France. L'objectif de cette première phase était de constituer le premier catalogue de gènes et de métagénomes du microbiote cæcal de poule en représentant une grande diversité des modes d'élevage. Ainsi, 340 échantillons dont nos 20 échantillons des lignées R+ et R- nourries avec l'aliment CTR (10 pour chacune des deux lignées), ont été séquencés (librairies pairées de 2x150pb sur un séquenceur Illumina HiSeq 4000) à grande profondeur (~80 millions de paires de séquences). Ils ont permis la construction d'un catalogue de référence que j'appellerai « catalogue MetaChick » (Plaza Oñate et al. 2023). Ce catalogue contient 13,6 millions de gènes, prédits pour coder des protéines et 2 629 MGS composés à 99,5% de bactéries.

La phase 2 du projet a débuté en 2021 et a consisté à utiliser ce catalogue de gènes et de métagénomes *via* le séquençage de nouveaux échantillons pour étudier le microbiote cæcal de poules dans conditions variées: analyse de l'influence de sélections divergentes variées, étude des flux de souches bactériennes et de gènes d'antibiorésistance, évaluation de l'impact de stress abiotiques comme un stress de chaleur (les échantillons HS du projet ChickStress) ou bien de l'impact d'une modification du régime alimentaire (nos échantillons LE). Cette deuxième phase de séquençage a été réalisée à plus faible profondeur (~40 millions de paires de séquences de 2x150 pb) sur un séquenceur Illumina Hiseq 4000.

b. Méthode et outils d'analyse des séquences de métagénomiques

Comme indiqué en introduction deux stratégies d'analyses de ces séquences sont possibles, l'analyse *de novo* avec une reconstruction d'un nouveau catalogue de référence en incluant dans notre cas le catalogue MetaChick pré-existant, ou l'analyse par la quantification des gènes et métagénomes du catalogue MetaChick. Nous avons choisi la première stratégie, et ce pour répondre à plusieurs objectifs : i) identifier de potentiels nouveaux gènes ou métagénomes présents dans nos échantillons LE. En effet, les paramètres environnementaux étant connus pour influencer la composition du microbiote, nous faisions l'hypothèse que ces nouveaux échantillons de poules pondeuses nourries avec un régime non conventionnel pourraient contenir de nouvelles espèces microbiennes. ii) Comparer l'analyse *de novo* seule aux résultats obtenus avec le séquençage métabarcoding en termes de reconstitution de la composition microbiennes et fonctionnelles[*].

Comme indiqué en introduction (paragraphe **A.5.2.a** et **Figure A-12**), l'analyse *de novo* des séquences de métagénomique se décompose en plusieurs étapes : 1) assemblage des lectures en contigs, 2) annotation structurale, 3) regroupement des contigs (« binning ») en espèces métagénomiques (MGS), 4) annotation taxonomique et fonctionnelle. Pour notre étude nous avons utilisé la chaîne de traitement metagWGS, développée par la plateforme Genotoul-bioinfo (INRAE Toulouse) (Fourquet et al. 2022). Le pipeline, tel qu'utilisé dans notre étude, inclut l'ensemble des étapes d'assemblage, d'identification et annotation fonctionnelle des gènes, de « binning » et d'annotation taxonomique des MGS (**Figure C-6**).

Pour l'assemblage (**Figure C-6-1**), j'ai choisi d'utiliser l'assembleur MetaSPAdes (Nurk et al. 2017). En comparaison à MEGAHIT (D. Li et al. 2016), il permet un assemblage de meilleure qualité, au prix de ressources computationnelles plus importantes (Vollmers, Wiegand, and Kaster 2017; C. Yang et al. 2021). Pour améliorer la qualité de l'analyse, seuls les contigs ayant une taille minimale de 1kb sont utilisés par la suite, car cette taille correspond à la taille moyenne d'un gène bactérien (calculée sur 5 089 génomes complets procaryote téléchargés de la base RefSeq en mai 2024). Après annotation structurale (**Figure C-6-2**), les gènes identifiés sur chaque échantillon ou groupe d'échantillons sont « clusterisés » au seuil admis de 95% d'identité pour réduire la redondance (**Figure C-6-4**). L'annotation fonctionnelle qui suit, repose sur la base de données de gènes orthologues eggNOG 5.0 (Huerta-Cepas et al. 2019) qui retourne différentes annotations telles que les catégories fonctionnelles COG, les ontologies de gènes GO (The Gene Ontology Consortium 2017), et KEGG (Kanehisa et al. 2017), ainsi que les modules, réactions, et voies métaboliques KEGG, les domaines protéiques de la base de données PFAM (Letunic and Bork 2018), et les enzymes dégradant les carbohydrates de la base de

[*] Objectif qui n'a pas pu être atteint dans les délais impartis à ce travail de thèse mais décrit dans les perspectives de valorisation paragraphe **G.2**.

données CAZy (Levasseur et al. 2013; Drula et al. 2022). Les gènes (en particulier ceux codant pour les protéines) sont également annotés taxonomiquement. Pour cela, ils sont alignés sur la base de référence de séquences protéiques NCBI-nr (téléchargée en octobre 2023) grâce à l'outil Diamond (Buchfink, Reuter, and Drost 2021) et une taxonomie consensus est calculée à partir des meilleurs alignements.

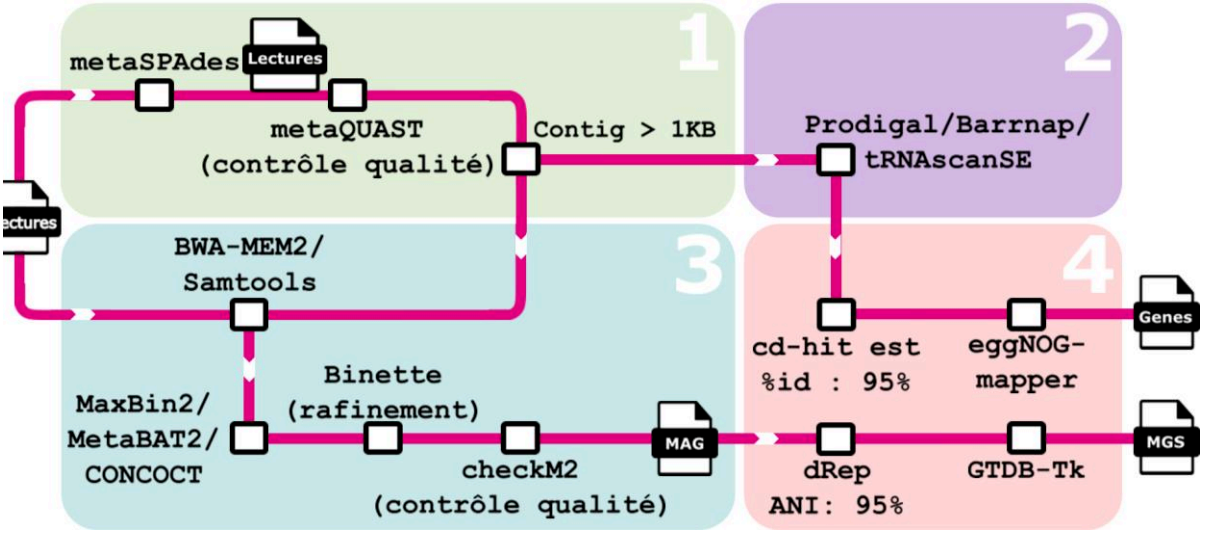


Figure C-6: Représentation graphique des outils et paramètres de metagWGS utilisés dans notre étude. Les briques correspondent à : 1. l'assemblage, 2. l'annotation structurale, 3. le « binning », 4. la réduction de la redondance entre échantillons ou groupes d'échantillons, et l'annotation fonctionnelle des gènes et taxonomique des MGS (figure adaptée de celle de la documentation de metagWGS).

Pour l'étape de reconstruction des MGS (**Figure C-6-3**), il existe une multitude d'outils qui associent les contigs principalement selon leurs compositions nucléotidiques, et leurs abondances (C. Yang et al. 2021). L'originalité de metagWGS est d'utiliser trois outils de « binning » et de combiner et d'affiner les résultats grâce à l'outil Binette (Mainguy and Hoede 2024). Les MAG ainsi obtenus sur chaque échantillon ou groupe d'échantillons sont ensuite dérépliqués au seuil ANI (pourcentage d'identité moyen) admis de 95% pour obtenir les MGS (**Figure C-6-4**). Ces derniers sont annotés taxonomiquement grâce à la base de données procaryotes GTDB version 214 (Parks et al. 2022).

Pour la quantification fonctionnelles (au travers de la quantification des gènes), j'ai choisi ici d'utiliser l'outil Meteor (Pons et al. 2010). Dédié à l'alignement de petites séquences, il permet d'aligner sur toute leur longueur les lectures sur les gènes d'un catalogue tout en contrôlant le pourcentage d'identité (ici fixé à 95%) entre les séquences. Par ailleurs, Meteor permet de quantifier les gènes (en nombre de lectures alignées) selon différentes méthodes, dont celle que j'ai choisi qui conserve les alignements multiples et redistribue leurs comptages en fonction des proportions d'alignements uniques sur les différents gènes ciblés. Le nombre de lectures alignées est ensuite normalisé par la

taille des lectures et la taille des gènes pour obtenir des abondances correspondant au nombre de copies de chaque gène. Ceci permet de pallier le fait que les gènes les plus longs ont tendance à avoir plus de comptage. Enfin, les abondances de ces gènes sont additionnées à l'échelle des fonctions (KEGG) pour réduire l'effet du nombre important d'abondance nulle ou très faible.

En ce qui concerne la quantification des taxonomies, comme indiqué en introduction, il existe également deux méthodes : par ré-alignement des lectures sur les MGS ou en se basant sur la quantification de gènes marqueurs. Pour tester la première méthode, j'ai utilisé l'outil BWA-MEM2 (Vasimuddin et al. 2019) puis j'ai calculé la profondeur de couverture de chaque MGS. Pour tester la seconde méthode, j'ai utilisé l'outils MSPminer (Plaza Oñate et al. 2019). Les gènes sont au préalable alignés sur les MAG (avec BLAST à 95% d'identité et 90% de couverture minimum), puis pour chaque MGS, l'abondance des gènes prévalents[*] dans au moins 50% des MAG qu'il représente, sont soumis à l'outils MSPminer pour en identifier les gènes marqueurs *i.e.* co-abondant. Pour chaque MGS et pour chaque échantillon ayant au moins 10% de gènes marqueurs avec une abondance non nulle, l'abondance moyenne des 100 gènes marqueurs qui corrèlent le plus est retenue comme l'abondance du MGS.

Les profils d'abondance des MGS et des fonctions sont ensuite analysés à partir des abondances raréfiées pour évaluer les indices de richesses, diversités et dissimilarités des groupes d'échantillons « lignée x régime ». L'analyse de ces indices a été faite grâce à un modèle linéaire incluant la lignée, le régime et, lorsque précisé, l'interaction de ces deux facteurs, soumis à une analyse de variance (package *car* (Fox et al. 2022)) suivie de test post-hoc (package *emmeans* (Lenth et al. 2022)). Pour identifier les MGS ou les fonctions différentiellement abondantes entre deux groupes d'échantillons, les abondances sont transformées en abondances relatives pour tenir compte des différences de profondeur de séquençage, puis normalisées par la méthode CLR (« Centered Log Ratio »). L'analyse d'abondance différentielle est réalisée grâce au package *Limma* (M. E. Ritchie et al. 2015a) qui implémente des statistiques modérées de Student à partir d'un modèle linéaire bayésien empirique (option `robust` activée). Les *P*-values sont ensuite ajustées par la méthode de test multiples de Benjamini and Hochberg, BH (Benjamini and Hochberg 1995). Le seuil de significativité des tests est fixé à 0,05.

L'analyse de ces séquences est décrite dans le chapitre **E**. Elle inclut une mise au point méthodologique qui compare les stratégies d'analyse *de novo*, les méthodes de quantification des espèces

[*] Prévalent : taux de présence dans un ensemble (ici nombre d'alignement d'un gène dans l'ensemble des MAG d'une même espèce représentée par une MGS.

métagénomiques, et inclut *in fine* une description du catalogue obtenu ainsi qu'une analyse des effets de la lignée et du régime sur la composition microbienne ou fonctionnelle.

C.3.3. Caractérisation du microbiote cæcal par séquençage complet de l'ARN

a. Acquisition des séquences de métatranscriptomique

Comme indiqué en introduction, cette troisième technique omique permettant la caractérisation du microbiote, est une technique plus récente et encore assez peu utilisée. Pour l'inclure dans cette étude, nous avons obtenu le financement du projet ChickMetaT sur un appel à projet INRAE du département de génétique animale en 2020. Ce projet s'est déroulé en deux temps, avec en premier lieu une mise au point des protocoles expérimentaux et dans un second temps l'application des procédures expérimentales sur l'échantillonnage complet incluant 60 échantillons, puis l'analyse des séquences conjointement aux analyses de séquences de métagénomique.

Bien que ma formation soit en bioinformatique, j'ai voulu suivre la mise au point des protocoles expérimentaux, en collaboration avec l'équipe GeMS de l'unité GABI. Ces phases essentielles ont eu pour but d'optimiser l'extraction des ARN à partir de contenus cæcaux, puis celle de la déplétion des ARN ribosomaux (la ribodéplétion), étape particulièrement délicate. Les échantillons stockés à -80°C depuis 2015 mais sans protection contre les ribonucléases (enzymes dégradant les ARN), ont été extraits par l'équipe GeMS avec le kit RNeasy PowerMicrobiome de QIAGEN dédié aux prélèvements intestinaux. Après comparaison avec le protocole classique d'extraction au trizol, ce kit permet une meilleure qualité et une plus grande quantité d'ARN extraits. Ensuite, il s'agit d'amplifier le signal des ARN messagers codant les protéines, en supprimant les ARNr qui représentent la quasi-totalité des ARN extraits. Le kit Ribo Zero, historiquement utilisé et produit par Illumina, a été retiré du marché en 2019, ce qui nous a obligé à mettre en place un comparatif d'autres kits pour choisir le protocole qui correspondait le mieux à nos besoins. Pour cela, j'ai donc contacté plusieurs laboratoires comme le Génoscope, la plateforme GenomEast (igbmc, Strasbourg) et des chercheurs INRAE pour recueillir des retours d'expérience. Après de nombreux échanges avec les commerciaux de différents fournisseurs, nous avons choisi de comparer 2 kits, QIAseq FastSelect (QIAGEN) et RiboPOOls (siTOOLs Biotech) puis de pratiquer un séquençage faible profondeur (sur un séquenceur Illumina MiSeq) pris en charge par la plateforme GeT-PlaGe (INRAE, Toulouse). Alors que le kit RiboPOOLs reprend le principe du kit historique, *i.e.* hybridation de sondes ciblant les ARNr, puis élimination des complexes sondes/ARNr grâce à des billes magnétiques, le kit FastSelect utilise des sondes ciblant les ARNr pour en bloquer la transcription inverse lors de la préparation des librairies de séquençage. A noter que le kit RiboPOOLs nécessite de sélectionner un mélange de sondes. Nous avons donc également testé deux solutions,

l'une contenant des sondes ciblant uniquement les bactéries (référencé « 100% bacteria » par la suite) et l'autre ciblant à 80% les ARNr bactériens, 10% les ARNr de champignons et 10% les ARNr de poule (référencé « Sondes mélangées » par la suite). Trois échantillons d'ARN totaux représentant trois niveaux de qualité d'ARN différents ont été utilisées pour comparer ces kits avec un échantillon répliqué pour chaque kit permettant de tester la reproductibilité de l'expérience. À la suite du séquençage des ARN ribodéplétés, un contrôle qualité de l'efficacité de la ribodéplétion a été réalisé *in silico* grâce à l'outils SortMeRNA (Kopylova, Noé, and Touzet 2012). Cet outil compare chaque lecture à une base de référence de séquences ribosomales (intégrant les bases Silva et Rfam (Kalvari et al. 2021)). Enfin, nous avons évalué le nombre de séquences finalement utiles, en mesurant le taux d'alignement des séquences non ribosomales sur le catalogue de gènes MetaChick.

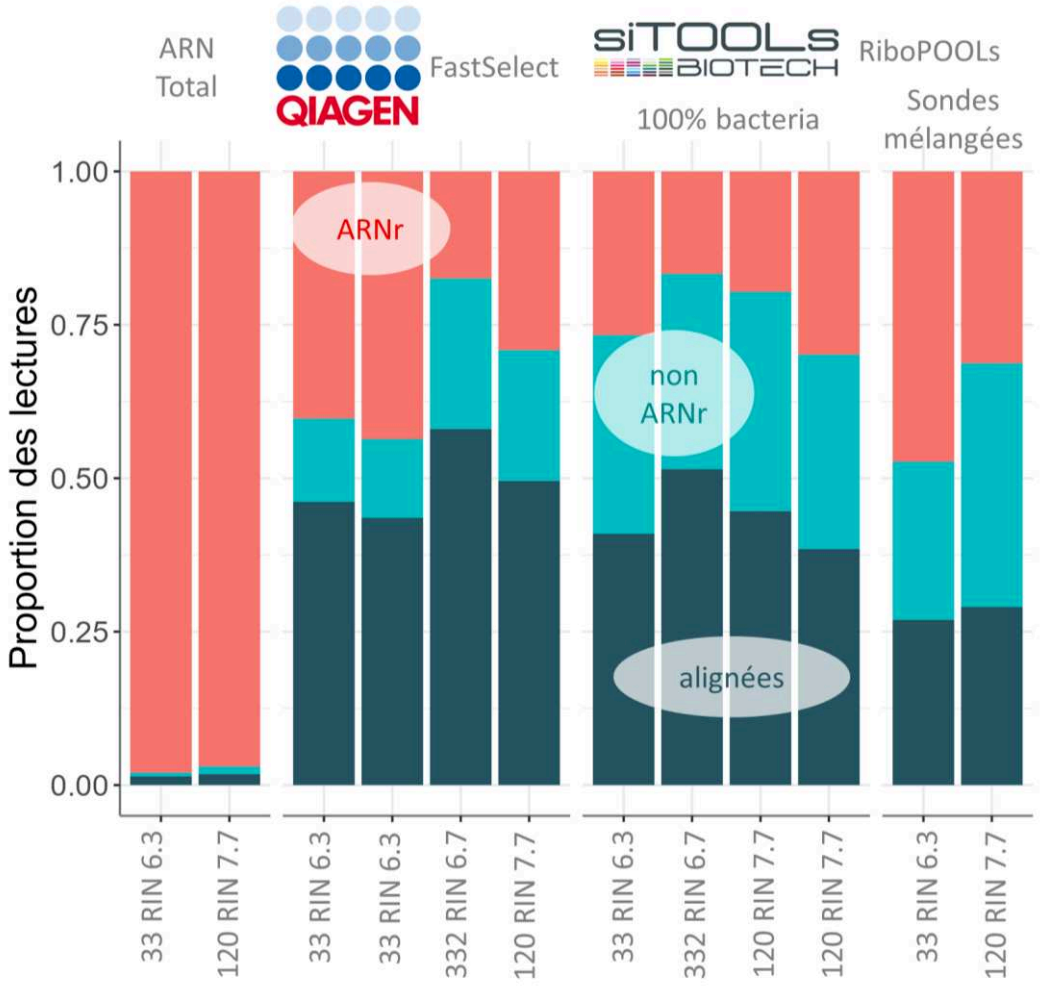


Figure C-7 : Proportion des lectures ribosomales, non ribosomales et alignées sur le catalogue MetaChick.
Les échantillons « ARN Total » correspondent aux échantillons non ribodéplétés. Les valeurs de RIN (« *RNA Integrated Number* »), reflètent la qualité des extractions des ARN totaux et s'étalent théoriquement sur une échelle de qualité ascendante de 1 à 10.

Ce projet pilote nous a permis de montrer l'efficacité des deux kits avec en moyenne 30% de lectures ribosomales restant contre plus de 97% dans les échantillons non ribodéplétés (**Figure C-7**). Pour le kit RiboPOOLs, l'utilisation de sondes 100% bactériennes permettait une meilleure ribodéplétion, ce qui est en cohérence avec la composition à très grande majorité bactérienne du microbiote cæcal. Le kit FastSelect semblait plus sensible à la qualité des ARN, mais plus reproductible et avec *in fine* un meilleur taux de réalignement des séquences sur le catalogue de gènes (moyenne à 73% au lieu de 57%). Finalement, pour les deux kits, nous conservions une proportion de lectures utiles similaires pour un coût similaire. Nous avons donc poursuivi l'analyse avec le kit RiboPOOLs ciblant uniquement les bactéries car la mise en place du protocole était plus simple. En effet, le produit de la ribodéplétion étant des ARN, ils peuvent être stockés et congelés en vue d'un séquençage ultérieur (dans notre étude sur un autre site, sur la plateforme GeT-PlaGe de Toulouse). Au contraire, les produits de la ribodéplétion avec le kit FastSelect sont des complexes sonde/ARN qui peuvent se désolidariser avec le temps. Les 41 échantillons ribodéplétés de ce projet de thèse ont ensuite été séquencés grâce à une librairie de séquençage pairé de 2x150 pb sur un séquenceur Illumina NovaSeq6000.

b. Méthode et outils d'analyse des séquences de métatranscriptomique.

L'analyse de ces séquences, tout comme en métagénomique, peut se réaliser *de novo*, avec la construction d'une référence de transcrits, ou bien en utilisant une référence connue.

Sur notre projet, nous disposons de séquences métagénomiques pour la plupart des échantillons séquencés ici en métatranscriptomique. Ces séquences, conjointement au catalogue de référence MetaChick, nous permette d'obtenir un catalogue enrichi de référence de notre écosystème cæcal. J'ai donc choisi de ne procéder qu'à une analyse quantitative de cette référence à partir des séquences de métatranscriptomique. Pour cela, les lectures provenant d'ARN ribosomaux sont identifiées et filtrées grâce à l'outils SortMeRNA. Les lectures non ribosomales sont ensuite alignées sur le génome de poule pour éliminer une potentielle contamination et enfin sur le catalogue enrichi de gènes avec l'outil Meteor avec le même paramétrage que précédemment (95% d'identité sur 100% de couverture). L'abondance de chaque gène tient compte des alignements uniques et multiples. Ces abondances, qui représentent ici l'expression des gènes, sont transformées en nombre de transcrits (normalisation par la taille des lectures et du gène), puis sommées à l'échelle des fonctions (KEGG). Les profils d'expression raréfiés permettent d'évaluer les richesses, diversités et dissimilarités des groupes d'échantillons (« lignée x régime ») grâce à des analyses de variance appliquée sur des modèles linéaires incluant la lignée, le régime et lorsque précisé, leur interaction. Les analyses d'expression différentielle entre groupe sont réalisées sur les abondances relatives normalisées CLR grâce au

package R Limma et un ajustement BH des *P*-values. Ces abondances normalisées sont également soumise à des analyses de redondance (RDA, package vegan (Oksanen et al. 2022)) utilisant des modèles incluant la lignée et/ou le régime. Enfin une analyse triadique partielle des données de métagénomique et de métatranscriptomique est également réalisée sur les abondances et expressions normalisées CLR pour évaluer leur ressemblance (package ade4 (Dray, Dufour, and Chessel 2007)). Le seuil de significativité des tests est fixé à 0,05.

Cette analyse quantitative des profils d'expression répond à deux objectifs qui sont présentés dans le chapitre **F** : i) évaluer l'apport d'informations issues de la métatranscriptomique par rapport à celles issues de la métagénomique, ii) identifier les fonctions différentiellement exprimées illustrant le rôle du microbiote intestinal dans l'adaptation à une modification du régime alimentaire ou dans l'efficience alimentaire des poules.

C.3.4. Les contrôles qualités préliminaires

A l'origine de cette thèse, le design expérimental incluait le design complet du projet ChickStress, *i.e.* avec les deux lignées supplémentaire Fayoumi et LS résistante à la chaleur, ainsi que la condition d'élevage simulant un stress de chaleur (HS). Les contrôles qualité, détaillés ci-après, ont été menés sur l'ensemble de l'échantillonnage (conditions « Control », « Low Energy » et « Heat Stress »), mais par souci de tenir le temps imparti à ces travaux dans le cadre d'une thèse, l'analyse proprement dite des séquences, s'est focalisée uniquement sur l'efficience alimentaire et l'impact d'une modification du régime.

a. Le métabarcoding ciblant l'ARNr 16S

Lors de la première analyse des données de séquences de métabarcoding ciblant le gène de l'ARNr 16S, les analyses classiques de richesse et de diversité ont mis en évidence une variabilité importante de composition du microbiote au sein de chaque groupe « lignée x régime », suggérant la présence de potentiels échantillons aberrants.

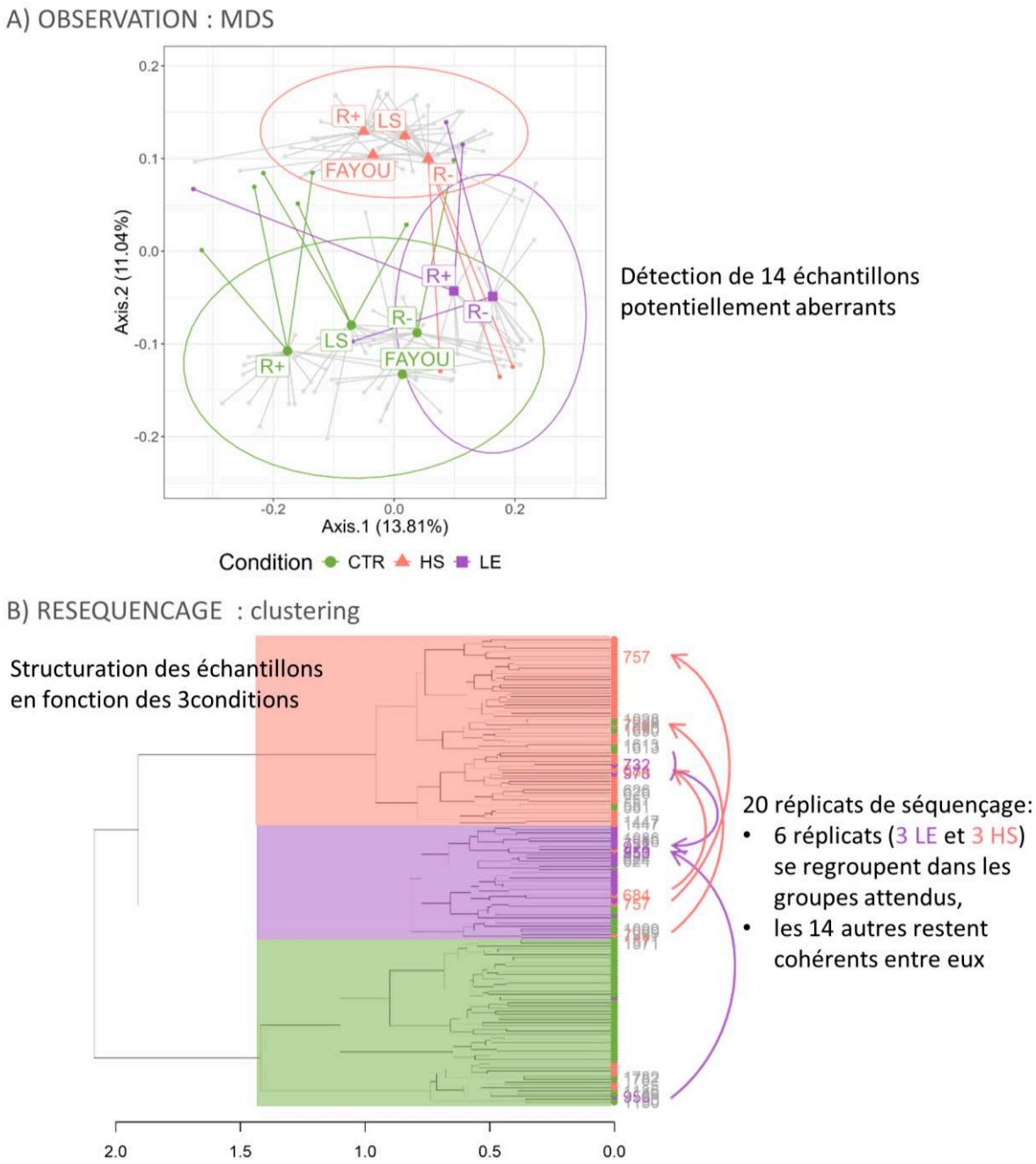


Figure C-8: Identification de potentiels échantillons aberrants et analyse de reproductibilité des échantillons re-séquencés.
A) Analyse en coordonnées principales (distance de Jaccard) mettant en évidence des échantillons hyper variables vis-à-vis de leur groupe (points colorés) ; B) *Clustering* hiérarchique (distance de Jaccard) de l'ensemble des séquençages. Les rectangles représentent la condition prédite à partir de ce clustering. Les noms des 20 échantillons séquencés deux fois sont indiqués. Les noms colorés représentent les réplicas pour lesquels le deuxième séquençage permet une meilleure classification de la condition d'élevage. Conditions CTR = contrôle, HS = stress thermique, LE = régime alimentaire appauvri.

Ainsi avant de poursuivre, nous avons décidé de procéder à un re-séquençage de 20 échantillons pour évaluer la reproductibilité de ces diversités et confirmer l'incohérence de certains échantillons. Le choix des échantillons à re-séquencer s'est fait en détectant :

- s'ils avaient un indice de diversité α (richesse, Chao1, Shannon ou inverse Simpson) extrêmes (1,5 fois la distance interquartile en dessous du 1er quartile ou au-dessus du 3e quartile) par rapport à leur groupe « lignée x régime »,
- si la prédiction de la condition d'élevage faite à partir des ordinations (analyse en coordonnées principales, PCoA) ou celle à partir des *clustering* hiérarchiques des échantillons basés sur 3 indices de dissimilarité β (Jaccard, Bray Curtis ou Unifrac) étaient systématiquement fausses.

Quatorze échantillons validaient deux de ces tests, 6 autres échantillons ont été sélectionnés à titre de contrôle (**Figure C-8-A**).

Après re-séquençage, une analyse des réplicas deux à deux a permis de valider une bonne reproductibilité des compositions microbiennes, avec pour la majorité des échantillons, une différence faible des indices de diversité α et β entre réplicas (différence/distance entre réplicas inférieure aux différences vis-à-vis de la médiane ou aux distances vis-à-vis du centroïde du groupe). L'analyse indique également une cohérence des prédictions de la condition d'élevage entre réplicas que ce soit sur les PCoA ou sur les *clustering* hiérarchiques (**Figure C-8-B**). Enfin elle a permis de valider le statut « d'outliers » de 6 échantillons (3 de la condition LE et 3 de la condition HS), dont le re-séquençage a permis de diminuer la variabilité intra-groupe et de prédire correctement la condition de l'échantillon. Pour la suite de l'analyse, le second séquençage a été utilisé pour les échantillons aberrants et celui avec le plus de profondeur a été conservé pour les autres échantillons répliqués.

b. La métagénomique, séquençage complet de l'ADN

L'analyse de données de séquences de métagénomique se divise en deux grandes étapes : la construction d'une référence de gènes et de métagénomes, puis leur quantification. Cette seconde étape permet également une validation qualitative du séquençage dans la mesure où la référence choisie est proche de l'écosystème étudié. Grâce à la phase 1 du projet MetaChick, nous avons bénéficié d'une référence du microbiote cæcal de poule incluant notamment nos échantillons contrôles, le catalogue MetaChick (voir paragraphe **C.3.2**). Pour contrôler la qualité du séquençage de la phase 2 du projet, les séquences issues des 60 échantillons relatifs aux conditions de modification du régime alimentaire, LE, et de stress thermique, HS, ainsi que les séquences de phase 1 des 40 échantillons contrôles, CTR, ont été alignés sur ce catalogue (**Figure C-9-A**).

Cette première analyse nous a permis de voir que les séquences des échantillons contrôles (CTR Phase 1) s'alignaient en moyenne à 82% ± 0,5 sur le catalogue, alors que le taux d'alignement des séquences des échantillons LE et HS était en moyenne plus faible et plus variable (74% ± 7,9). En effet, ils sont compris entre 60% et 83% et séparés en deux groupes selon qu'ils soient alignés avec un taux inférieur à 73% (50% des échantillons, référencés « low » par la suite) ou supérieur à 73% (50% des échantillons, référencés « high » par la suite).

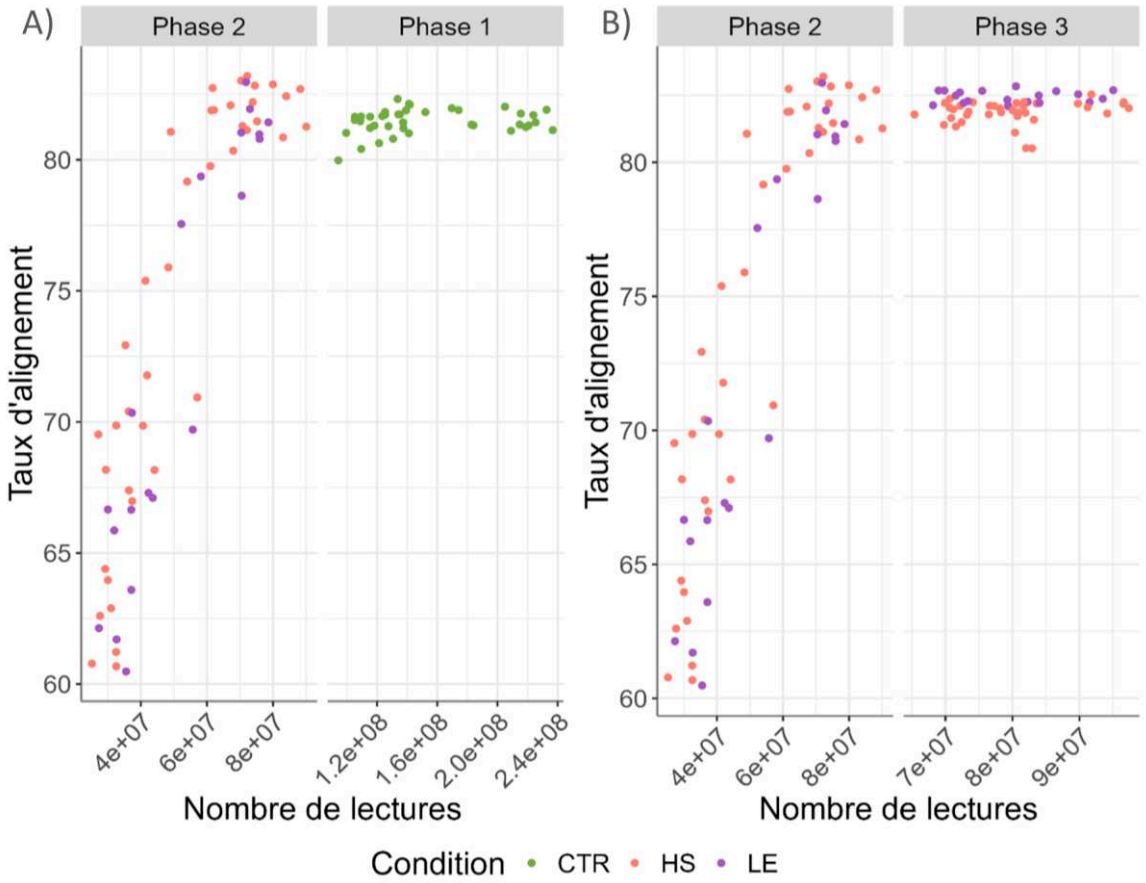


Figure C-9 : Distribution du nombre de lectures et du taux d'alignement sur le catalogue MetaChick.
A) Comparaison des échantillons stressés (HS, LE) séquencés en phase 2 et des échantillons contrôles (CTR) séquencés en phase 1 ; B) Comparaison des échantillons stressés séquencés en phase 2 puis re-séquencés en phase 3.
Conditions CTR = contrôle, HS = stress thermique, LE = régime alimentaire appauvri.

Ce faible taux d'alignement peut s'expliquer par des raisons biologiques, les échantillons « low » pourraient contenir de nouveaux gènes ou génomes qui seraient absents du catalogue MetaChick ou alors être le résultat d'un biais technique lors de la génération des séquences. Deux études ont été menées en parallèle, pour tenter de valider rapidement l'une de ces deux hypothèses.

J'ai d'abord procédé à une analyse *de novo* des lectures non alignées sur le catalogue, pour chacun des groupes (contrôle, high, low), autrement dit une assignation taxonomique des lectures ainsi que des assemblages de ces lectures. Cette analyse n'a pas permis de mettre en évidence une réalité biologique de ce biais d'alignement. En revanche, l'analyse des contrôles qualités et des plans d'expériences des protocoles expérimentaux ont mis en évidence des différences entre les échantillons « high » et « low ». En effet, les échantillons « low » après extraction et après construction des librairies de séquençage étaient en moyenne moins concentrés en ADN, et malgré une normalisation de ces concentrations avant séquençage, ont en moyenne moins de séquences (17 contre 34 millions de paires pour les échantillons « high »). Grâce au plan d'expérience, nous avons pu identifier que les ADN de l'ensemble des échantillons « low » ont été extraits ensemble sur des plaques différentes de ceux des échantillons « high ». Ces extractions ont été réalisées en partie manuellement et en partie *via* un robot, mais les contrôles de concentration faits par densité optique n'ont à l'époque pas permis de mettre en évidence ce biais. Malheureusement, nous n'avons pu poursuivre davantage l'identification de la cause de ce biais.

Afin de corriger cet effet plaque, l'ensemble des échantillons stressés a été extraits à nouveau en utilisant un autre protocole d'extraction (similaire à celui utilisé pour les échantillons contrôles), puis re-séquencés et réalignés sur le catalogue MetaChick (**Figure C-9-B**). Les séquences des échantillons de cette 3^{e} phase de séquençage présentent cette fois-ci, un taux d'alignement homogène en moyenne de 82% ± 0,5, comparable à celui des échantillons contrôles de phase 1. Par ailleurs, ce nouveau séquençage nous a permis d'obtenir une plus grande profondeur de séquençage avec en moyenne 40 millions de paires de séquences par échantillon.

L'ensemble de ces contrôles qualité a fait partie du travail réalisé pendant ma première année de thèse. Ils ont fait l'objet d'un poster présenté lors du séminaire des doctorants du département de Génétique Animale de 2022 (Bernard et al. 2022) :

Bernard, M., Coville, J-L., Bruneau, N., Jardet, D., Calenge, F., Pascal, G., Zerjal, T. PERFORM PROTOCOL TESTS AND SEQUENCE CHECKS ? THE EXPERIENCE FROM MULTIOMICS MICROBIOTA DATA. *25ème Séminaire des Doctorants du Département de Génétique Animale (SDDGA 2022)*, Sep 2022, Bordeaux, France. https://hal.science/hal-03930031

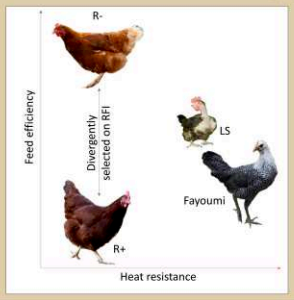
WHY PERFORM PROTOCOL TESTS AND SEQUENCE CHECKS ? THE EXPERIENCE FROM MULTI-OMICS MICROBIOTA DATA

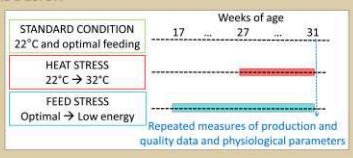
Maria Bernard[1], Jean-Luc Coville[1], Nicolas Bruneau[1], Déborah Jardet[1], Fanny Calenge[1], Géraldine Pascal[2], Tatiana Zerjal[1]
[1] : Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France; [2] : GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326 Castanet Tolosan, France

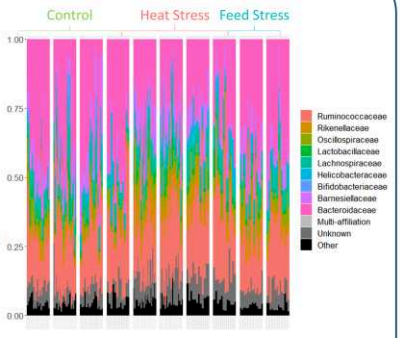
How to differenciate high microbiota natural variability from technical bias?

METABARCODING

EXPERIMENTAL DESIGN



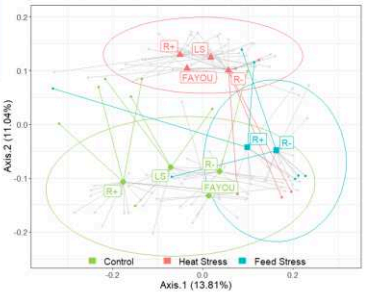




16S sequences have been analysed with FROGS (1), and bacterial composition caracterised with Phyloseq (2) R package through:

- Family composition profiles
- 4 α-diversity indices
- 3 β-dissimilarity methods used for MDS and hierarchical clustering

OBSERVATION : MDS on Jaccard distance



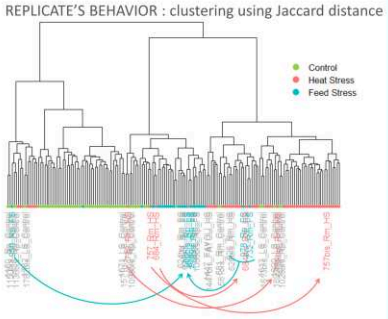
→ Should the 14 suspicious samples be considered as samples with high natural variability or outliers ?

Resequencing of 20 samples allowed us to:

- Validate that **sequencing is quite well reproducible** even if the composition profile still shows variability, and **confirm the high natural variability**
- **Identify 6 outliers** samples that need to be removed

→ **Next step is to select one sample per animal by keeping the closest sample to the average profile of its group or with the deepest sequencing.**

REPLICATE'S BEHAVIOR : clustering using Jaccard distance

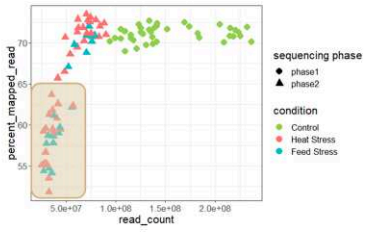


How to choose the right protocols for metagenomics shotgun and metatranscriptomics ?

METAGENOMICS

Metagenomic samples were produced in two batches. Sequences have been quality checked and aligned on the internal gene chicken gut catalog using meteor (3).

→ Is the low mapping rate due to biological reason or technical bias ?



- No obvious biological reasons were discovered by unmmapped read assembly and taxonomical affiliation.
- History of sample experiments highlighted a technical bias in the second phase DNA extraction protocol.

→ **Reextraction and resequencing of all stress samples as done in phase1 are in progress.**

METATRANSCRIPTOMICS

Metatranscriptomics is still an emerging omics in gut microbiota study, with a delicate protocol step, the ribodepletion.

→ Which ribodepletion kit is most suitable for our chicken ceacum samples?

Remaining rRNA were identified through SortMeRNA (4), kits were evaluated on their ease of use, cost, efficiency, sensitivity, and reproducibility.

Setting up a ribodepletion kit test project :

- 3 kits : Illumina, Qiagen, SiTools (2 probe mixes)
- 3 samples of different RIN Scores
- 1 constraint : ribodepletion experiment at Jouy en Josas and sequencing at getPlage Toulouse.

- Illumina : unusable in our experimental context, very expensive
- SiTools (100% bacteria) vs QIAGEN :
  - equivalent cost and efficiency
  - probably a better reproducibility for QIAGEN but a higher RIN score sensitivity
  - in our experimental context, a greater ease of use for SiTools

→ **Sitools 100% bacteria was retained and sequences on the project samples were just received.**

References

1 : F.Escudié et al., 2018, FROGS: Find, Rapidly, OTUs with Galaxy Solution, Bioinformatics, Volume 34, Issue 8, Pages 1287–1294
2 : McMurdie and Holmes, 2013, phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE. 8(4):e61217
3 : Pons N et al., 2010, METEOR, a platform for quantitative metagenomic profiling of complex ecosystems. JOBIM 2010 conference.
4 : Evguenia K. et al. 2012, SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data, Bioinformatics, Volume 28, Issue 24, Pages 3211–3217

Fundings

ANR project Chickstress, for animal production; European project Feed-a-Gene, for metabarcoding sequencing; France Génomique project MetaChick for metagenomic sequencing and INRAE AG department project ChickMetaT for metatranscriptomic.

Acknowledgements

The authors are grateful to the Genotoul bioinformatics platform Occitanie (doi: 10.15454/1.5572369328961167E12) and Migale bioinformatics facility (doi: 10.15454/1.5572390655343293E12) for providing help, computing and storage resources. Samples have been extracted by INRAE @Bridge and SAMBO plateforms, and sequences have been produced by INRAE Get-Plage and Genoscope plateforms.

XXVème SDDGA
19-20 mai 2022

INRAE

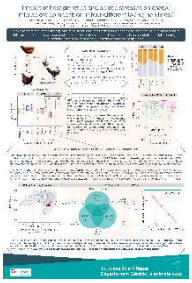
Poster présenté lors du séminaire des doctorants du département de Génétique Animale 2022.

D. PARTIE 1 : L'ANALYSE DE DONNEES DE METABARCODING REVELE DES INTERACTIONS HOTES-MICROBIOTES DEPENDANTES DU REGIME

L'objectif de cette première partie de résultats est de répondre aux deux premières questions de la thèse en décrivant, grâce aux données de séquençage de métabarcoding, les compositions microbiennes et les éventuelles fonctions associées présentes dans le cæcum de poule pondeuse. Cette première analyse tente de mettre en évidence les liens potentiels entre le microbiote et l'efficience alimentaire des poules et/ou le régime alimentaire utilisé. Elle a été valorisée par plusieurs communications scientifiques (paragraphe **D.5**) :

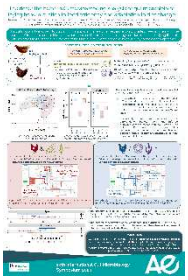
- Un poster présenté en séminaire interne :

Bernard, M., Coville, J.-L., Bruneau, N., Jardet, D., Lagarrigue, S., Calenge, F., Pascal, G., & Zerjal, T. Impact of host genetics and abiotic stresses on cæcal microbiota composition in four different laying hen lines? Journées Scientifiques du Département de Génétique Animale, September 2022 à Bordeaux, France. https://hal.inrae.fr/hal-03930049 :



- Un poster présenté lors d'un congrès international :

Bernard, M., Lecoeur, A., Coville, J.-L., Bruneau, N., Jardet, D., Lagarrigue, S., Troegeler-Meynadier, A., Calenge, F., Pascal, G., & Zerjal, T. The richer the better: 16S metabarcoding analysis of gut microbiota of laying hens in relation to feed efficiency and adaptation to diet change. International Gut Microbiology Symposium, Jun 2023 at Aberdeen, Scotland. https://hal.inrae.fr/hal-04127869 :



- D'une communication orale lors d'un congrès international :

Bernard, M., Lecoeur, A., Coville, J.-L., Bruneau, N., Jardet, D., Lagarrigue, S., Calenge, F., Pascal, G., & Zerjal, T. A richer gut microbiota is related to better feed efficiency and diet adaptability in laying hens. *74. Annual meeting of the European federation of animal science (EAAP)*, August 2023 at Lyon, France. https://hal.inrae.fr/hal-04199104

- Un article publié dans Scientific Reports dont le contexte, la démarche et les résultats sont résumés dans les paragraphes qui suivent.

Bernard, M., Lecoeur, A., Coville, J.-L., Bruneau, N., Jardet, D., Lagarrigue, S., Meynadier, A., Calenge, F., Pascal, G., & Zerjal, T. (2024). Relationship between feed efficiency and gut microbiota in laying chickens under contrasting feeding conditions. Scientific Reports, 14(1), 8210. https://doi.org/10.1038/s41598-024-58374-3 :


scientific reports

OPEN Relationship between feed efficiency and gut microbiota in laying chickens under contrasting feeding conditions

D.1. Contexte et objectifs

Les œufs de poules sont consommés partout dans le monde et représentent la ressource en protéine pour l'alimentation humaine la moins coûteuse. Pour répondre à la consommation croissante d'œufs, les entreprises de sélection avicole ont sélectionné des animaux de plus en plus efficients. Ces animaux sont ensuite exportés partout dans le monde et donc élevés dans des environnements géographiques et climatiques, conditions d'élevage mais également régimes alimentaires plus diversifiés que ceux dans lequel ils ont été sélectionnés. L'efficience alimentaire, un des phénotypes majeurs sous sélection chez les animaux d'élevage dont la poule, est sous le contrôle d'une multitude de facteurs notamment génétiques et environnementaux. Ces dernières années, les analyses sur le microbiote intestinal de différentes espèces animales d'élevage, suggèrent qu'il pourrait également jouer un rôle dans l'efficience alimentaire des animaux. En effet, son rôle dans la dégradation des aliments non digérés par les enzymes de son hôte et la production de métabolites qui en découlent, seraient directement bénéfiques pour le métabolisme notamment énergétique des animaux. Ainsi l'énergie apportée par le

microbiote intestinal représente autour de 10% chez l'homme, la poule ou le porc (McNeil 1984; Józefiak, Rutkowski, and Martin 2004; Bai et al. 2022), 40% chez le lapin (Marty 1984) et plus de 70% chez les ruminants (Liu et al. 2022). Les connaissances sur le microbiote intestinal de poule augmentent mais les résultats sur son association avec l'efficience alimentaire ne sont pas consensuels. Par ailleurs, les analyses se concentrent majoritairement sur les poulets de chair. Or bien que de la même espèce, les poulets de chair diffèrent génétiquement des poules pondeuse car ils ne sont pas sélectionnés sur les mêmes caractères. Ils sont sélectionnés pour une croissance rapide et une prise de masse musculaire importante, tandis que les poules pondeuses sont sélectionnées pour leur capacité à produire des œufs de manière efficace. Des différences existent également dans l'âge et le sexe des animaux utilisés dans les études, ce qui reflète la réalité des animaux d'élevage : les poulets de chair sont souvent des jeunes, mâles et femelles, tandis que les poules pondeuses sont des femelles adultes. Enfin, du fait de l'ensemble de ces différences, les animaux ne sont en général pas nourris avec le même le régime alimentaire. Or tous ces facteurs sont connus pour avoir des effets plus ou moins importants sur la composition du microbiote.

C'est dans ce contexte que se positionne cette première étude. En utilisant les lignées R+ et R- divergentes pour l'efficience alimentaire et soumises à deux régimes alimentaires, l'un riche en amidon et l'autre appauvri en énergie (-15%) et riche en fibres, elle répond à deux objectifs : 1) le microbiote contribue-t-il à l'efficience alimentaire des poules pondeuses adultes ? 2) Est-ce que cette contribution est conditionnée par le régime alimentaire ? L'étude repose sur l'utilisation de séquences issues d'un séquençage métabarcoding de l'ARNr 16S permettant l'identification des communautés bactériennes et une inférence des fonctions portées par celles-ci.

D.2. Matériels et méthodes

Pour rappel du design expérimental de cette étude (détaillé au paragraphe **C.2**), 58 poules de deux lignées divergentes pour l'efficience alimentaire, R+ (non efficiente) et R- (efficiente) sont nées au même moment et ont été élevées ensemble jusqu'à l'âge de 17 semaines. Elles ont ensuite été placées en cage individuelle pendant 14 semaines et nourries *ad libitum* avec l'un des deux régimes : le régime contrôle (CTR, 18 poules par lignée) est riche en amidon et sa composition est dominée par le blé et le soja, et le régime appauvri de 15% en énergie (LE, 11 poules par lignées) contient notamment -23% d'amidon, 2,1 fois plus de cellulose et une diversité de céréales plus importantes que le régime CTR (Tableau 4 de l'article publié dans Scientific Reports, paragraphe **D.5**). Les contenus cæcaux ont été prélevés à 31 semaines, et l'ADN du microbiote a été séquencé *via* la technique de métabarcoding ciblant la région V3-V4 du gène de l'ARNr 16S.

L'analyse bioinformatique des séquences a été réalisée comme indiqué précédemment (paragraphe **C.3.1.b**) avec la suite logicielle FROGS permettant la construction d'ASV annotés taxonomiquement, suivie par une analyse par inférence fonctionnelle de ces ASV. L'interprétation des résultats repose sur les principes classiques en écologie, notamment l'analyse de richesse et de diversité de chaque communauté par la mesure de l'alpha diversité, et l'analyse des dissimilarités entre communautés par le calcul de la béta diversité. Par ailleurs pour identifier les acteurs microbiens clés caractérisant les effets combinés ou non des deux variables de notre plan expérimental, *i.e.* la lignée et le régime, des analyses d'abondances différentielles ont été réalisées. Ces analyses statistiques ont été conduites sous R en utilisant majoritairement les packages dédiés à la description des communautés d'un écosystème, Phyloseq et vegan, ainsi que le package dédié à l'analyse d'abondance différentielle DESeq2.

Grâce à l'analyse des spécificités taxonomiques (Figure 3 de l'article, paragraphe **D.5**) et fonctionnelles (Figure 4 de l'article, paragraphe **D.5**) des différents microbiotes, nous avons pu émettre l'hypothèse d'une production différenciée d'acide gras à chaîne courte (*« short chain fatty acid »*, SCFA), dont les trois principaux sont l'acétate, le propionate et le butyrate, en fonction de la lignée et/ou du régime alimentaire. En effet, l'analyse d'abondance différentielle des fonctions des métabolismes du propionate et du butyrate, suggère une production accrue de ces SCFA chez la lignée R- vis-à-vis de la lignée R+ lorsqu'elles étaient nourries avec le régime CTR, ainsi qu'une production accrue du butyrate chez la lignée R+ lorsqu'elles sont nourries avec le régime LE en comparaison du régime CTR. Pour valider cette hypothèse, nous avons collaboré avec l'unité Micalis (INRAE, Jouy-en-Josas) pour nous former et réaliser un dosage de ces acides gras par chromatographie gazeuse. Les analyses comparatives des productions de SCFA ont été réalisées en proportion relative des trois acides gras majoritaires qui composent plus de 97% des acides gras produits. En effet, les mesures brutes de quantité de ces acides gras peuvent être influencées par une différence de production globale ou bien par une différence de quantité déjà absorbée.

D.3. Travaux préliminaires à l'analyse d'inférence fonctionnelle

L'analyse fonctionnelle réalisée à partir de données de séquençage amplicon ne peut se faire que par inférence des fonctions présentes. Dans cette analyse, nous avons utilisé les outils FROGS basés sur la suite PICRUSt2. La première étape de cette analyse est une étape clé, puisqu'elle recherche le plus proche parent phylogénétique et annoté fonctionnellement de chaque ASV, pour y associer tout ou partie des fonctions de ce parent. Pour identifier ce plus proche parent, PICRUSt2 utilise un arbre phylogénétique de séquences du gène de l'ARNr 16S préalablement construit, et positionne chaque

séquence d'ASV dans cet arbre. Pour évaluer la proximité de chaque ASV avec son plus proche parent, PICRUSt2 fournit un index, le NSTI (*« Nearest Sequence Taxon Index »*) représentant la somme des longueurs de branches qui séparent les deux séquences. Plus l'index est faible plus les séquences sont proches. Il est généralement admis qu'une distance supérieure à 2 (valeur par défaut de filtre dans PICRUSt2) correspond à du bruit qu'il faut exclure, qu'une distance entre 1 et 2 est médiocre, entre 0,5 et 1 est moyenne et qu'une distance inférieure à 0,5 est bonne. Un seuil de 0,5 ne signifie pas que les deux séquences comparées sont de la même espèce, mais considérant la redondance des fonctions entre micro-organismes, il n'est pas nécessairement utile d'avoir une correspondance exacte pour en tirer des informations fonctionnelles intéressantes. Pour autant, certaines fonctions sont spécifiques d'une petite communauté de micro-organismes, voire spécifiques d'une espèce. Dans ce dernier cas, la proximité avec la séquence de référence est primordiale. Comme pour beaucoup de filtres, il faut donc trouver un équilibre entre d'une part la représentativité des communautés prises en compte et donc leur abondance, ce qui va influencer la quantification des fonctions, et d'autre part la précision des prédictions fonctionnelles, ce qui va influencer la sensibilité et la spécificité de l'inférence. L'approche mise en place est ici de fixer nos paramètres afin de conserver les prédictions fonctionnelles des espèces en minimisant le NSTI en dessous de 0,5 et en maximisant la proportion des abondances prises en compte de ces mêmes espèces. En suivant cette démarche, le seuil maximal de NSTI a été fixé à 0,31 pour cette analyse (**Figure D-1-A**).

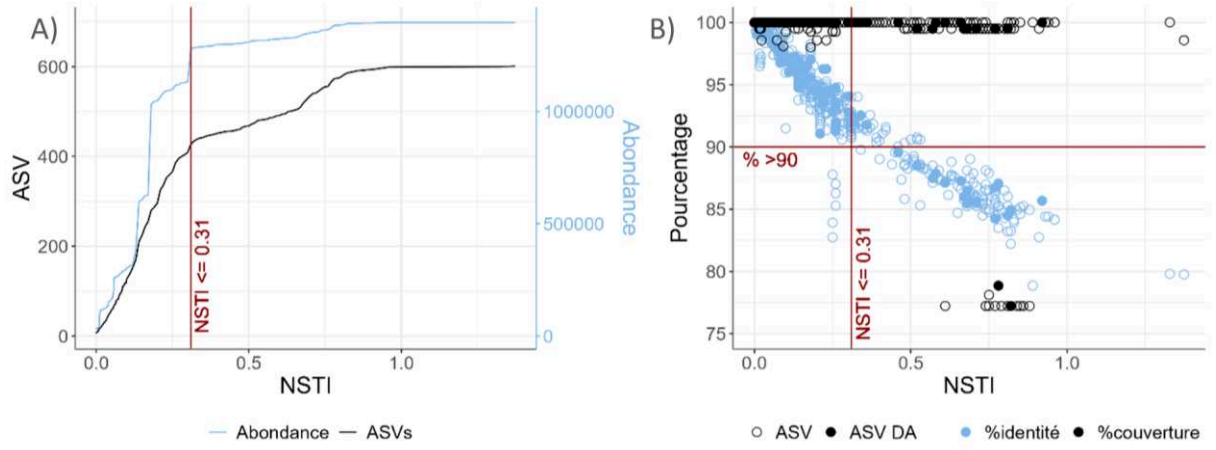


Figure D-1 : Distribution des valeurs de NSTI en fonction du pourcentage d'identité et de la couverture de l'alignement.
A) Nombre d'ASV et abondances des séquences associées en fonction d'une valeur minimum de NSTI; B) Distribution des ASV différentiellement abondant (ASV DA) ou non (ASV) en fonction de leur valeur NSTI, de leur pourcentage d'identité et de couverture avec la séquence du plus proche parent. Cette figure correspond à la figure S3 de l'article publié dans Scientific Reports (paragraphe **D.5**).

Pour aller plus loin dans la caractérisation de la proximité entre les ASV et les séquences de référence de PICRUSt2, nous avons aligné chaque séquence d'ASV avec la séquence du plus proche parent de PICRUSt2 et relevé ainsi les mesures de pourcentage d'identité et de couverture entre ces deux séquences (**Figure D-1-B**). Ainsi, avec un NSTI inférieur à 0,31 et en appliquant un seuil de 90% d'identité sur au moins 90% de couverture, nous avons retenu 421 ASVs (soit 70% des ASVs correspondant à 91% des abondances). Ceci signifie, que les inférences fonctionnelles déduites après analyse par FROGSFUNC porteront donc sur 91% des séquences bactériennes initiales de notre jeu de données ce qui est une représentativité tout à fait correcte. Cette procédure a été valorisée au cours de la thèse par la création d'une nouvelle version de la suite FROGS dont j'ai supervisé le développement (4.1.0 en mai 2023, https://github.com/geraldinepascal/FROGS/releases/tag/v4.1.0).

D.4. Résultats et Discussion

En préambule, il faut noter que notre échantillonnage respecte tout d'abord les caractéristiques physiologiques et de performances observées à l'échelle des populations plus grandes des lignées R+ et R-. Ainsi la lignée efficiente R- montre des mesures d'efficience nettement améliorées vis-à-vis de la lignée non efficiente R+ qui mange près de 1,5 fois plus (Tableau 1 de l'article, paragraphe **D.5**), mais ne diffère ni sur le poids des poules, ni sur la ponte. Par ailleurs, ces caractéristiques sont maintenues quel que soit le régime alimentaire. Il faut également noter que le régime alimentaire appauvri en énergie (LE) impacte négativement l'efficience des poules et ce quelle que soit la lignée (elles compensent la perte d'énergie en augmentant la quantité d'aliment ingéré sans impacter les caractères de ponte).

De façon très intéressante, cette première analyse du microbiote cæcal de poule a montré un motif récurrent de structuration de nos quatre groupes qui est illustré par la Figure 1 de l'article. La lignée non efficiente R+ nourrie avec le régime CTR (R+ CTR) présente un microbiote nettement différent de ceux des deux lignées nourries avec le régime LE (R- LE et R+ LE) qui ne présentent pour ainsi dire aucune différence entre eux. La lignée efficiente R- nourrie avec le régime CTR (R- CTR) présente un microbiote intermédiaire entre ces deux groupes et partage beaucoup de caractéristiques des microbiotes obtenus avec le régime LE. Ces résultats montrent que les différences que nous pourrions attribuer à la génétique de l'hôte (différence entre lignée) sont modulées en fonction du régime alimentaire. Cette interaction est notamment caractérisée par une richesse microbienne beaucoup plus faible chez les poules non efficientes R+ CTR que chez les poules efficientes R- CTR (+26%) ou que chez les poules nourries avec le régime LE (+36%) (Figure1-a de l'article, paragraphe **D.5**). Elle se caractérise également par des ASV et des fonctions différentiellement abondantes lorsque nous

comparons les lignées nourries avec le régime CTR, mais presque plus sous le régime LE. De même, moins de différences provoquées par le changement de régime sont identifiées chez la lignée efficiente R- que chez la lignée non efficiente R+. Les analyses taxonomiques et fonctionnelles qui découlent de ces observations nous ont menés à contrôler la production de SCFA produits par le microbiote. Cette expérience de dosage a également confirmé l'interaction entre la lignée génétique et le régime alimentaire par des proportions différenciées de ces acides gras (Tableau 3 et Figure supplémentaire S1 de l'article, paragraphe **D.5**) : la proportion de proprionate produit est supérieure chez R- CTR et chez les deux lignées nourries avec le régime LE ; et la proportion de butyrate tend à être supérieure chez R+ CTR comparé aux trois autres groupes. Pour expliquer ces différences nous avons émis l'hypothèse que la composition du digesta atteignant le cæcum diffère dans ces groupes.

En effet, cette différence est logiquement attendue lorsque l'on compare les deux régimes dont les compositions sont très différentes, et entraîne des contraintes de digestibilité différentes. Lorsque l'on compare les deux lignées, la différence de substrat entrant dans le cæcum pourrait s'expliquer par une différence de capacité digestive en fonction de la quantité d'aliment ingérée. Par exemple, la digestibilité de l'amidon (présent en forte proportion dans le régime CTR) est globalement bonne chez la poule mais jusqu'à une certaine quantité. Ainsi, pour la lignée R+ nourrie avec le régime CTR, la quantité d'aliment ingérée pourrait amener à une plus grande proportion d'amidon résistant dans le digesta cæcal. Cette quantité accrue d'amidon pourrait expliquer la surabondance observée d'Actinobacteriota et en particulier du genre *Bifidobacterium* dans le microbiote des poules R+ CTR. Elle pourrait également expliquer la surabondance dans ce groupe de fonctions impliquées dans le métabolisme de l'amidon et du saccharose et enfin la proportion plus élevée de butyrate, le SCFA favorisé lors de la fermentation de l'amidon. Au contraire, sous le régime LE, la différence de quantité d'aliment ingérée est toujours présente entre les deux lignées, mais la digestibilité de ce régime étant globalement difficile, nous faisons l'hypothèse que la composition du digesta est similaire peu importe la quantité ingérée. Ainsi pour les deux lignées nourries avec le régime LE, cette similarité de composition de digesta favoriserait une composition taxonomique et fonctionnelle du microbiote faiblement différenciée. En l'occurrence, la proportion accrue de fibres indigestes (en particulier la cellulose) pourrait être responsable de la surabondance de bactéries fibrolytiques du phylum Bacteroidota, bactéries classées comme productrices de propionate. Enfin, pour tenter d'expliquer le statut intermédiaire du microbiote R- CTR, nous pouvons combiner les deux hypothèses précédentes : avec une quantité d'aliment ingérée moindre que chez la lignée R+, la digestibilité de l'amidon dans la partie haute de l'intestin ne doit pas être problématique favorisant une composition du digesta cæcal autour des fibres indigestes également présentes dans le régime CTR. La composition des communautés microbiennes serait alors favorisée vers la dégradation de ces fibres et vers une

production accrue de propionate. Le fait que des différences persistent entre le microbiote R- CTR et les microbiotes associés au régime LE peut s'expliquer par une variation des sources de fibres incluses dans chacun des régimes. A noter que, le propionate produit en plus grande proportion dans ces trois groupes (R- CTR, R+LE et R- LE) est un acide gras qui passe la barrière intestinale et atteint le foie dans lequel il va stimuler la néoglucogénèse. Par ailleurs, des analyses de transcriptomique du foie ont montré une surexpression des gènes clefs de la glucogénèse chez les poules efficientes R- CTR comparée aux poules non efficiente R+ CTR (données non publiées et non inclues dans ce projet de thèse).

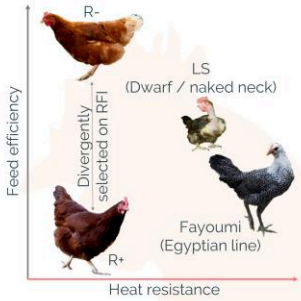
Pour conclure, notre étude montre que la sélection divergente sur l'efficience alimentaire ayant abouti à la création des lignées non efficiente R+ et efficiente R-, a provoqué de manière directe ou indirecte une modification forte de la composition microbienne. Pour autant, cette différenciation est dépendante du régime alimentaire, ce qui module l'association possible entre microbiote et efficience alimentaire. Dans le cas du régime CTR, le microbiote pourrait contribuer à une meilleure efficience alimentaire de la lignée R- par une capacité accrue à dégrader les fibres et une production plus importante de propionate.

D.5. Valorisations scientifiques

Impact of host genetics and abiotic stresses on caecal microbiota composition in four different laying hen lines?

Bernard M.[1], Coville J-L.[1], Bruneau N.[1], Jardet D.[1], Lagarrigue S.[2], Calenge F.[1], Pascal G.[3], Zerjal T.[1]
1 : Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350 Jouy-en-Josas, France; 2 : 1INRAE, INSTITUT AGRO, PEGASE, 35590 Saint-Gilles;
3 : GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326 Castanet Tolosan, France

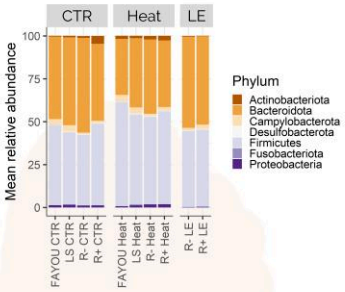
In the context of climate change and feed-food resources competition, the description of the gut microbiota of different chicken lines with distinct levels of feed efficiency and heat sensitivity, is a first step in determining its functional role in feed efficiency and environmental adaptation.

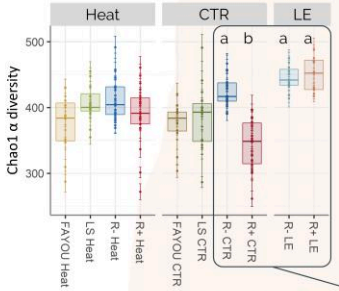


EXPERIMENTAL DESIGN

Four experimental chicken lines submitted to 3 rearing environments:
i) standard temperature and optimal diet, CTR;
ii) chronic heat stress (4 weeks at 32°C), Heat;
iii) chronic feed stress (from 17 to 31 weeks of age fed with a lower energy diet), LE.

The gut microbiota of 146 laying hens (from 11 to 20 samples per line and condition) were sequenced by 16S metabarcoding.

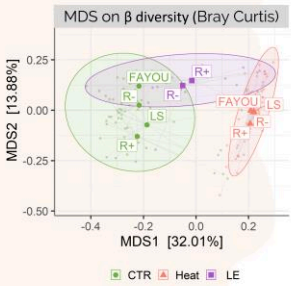




Anova p-values < 0,001 for line, condition and interaction

The alpha diversity measures showed **large differences among lines and conditions** with the particular case of the **non efficient line (R+) in CTR condition having an extremely low richness** in comparison with all other groups.

Beta diversity analysis confirmed the **line effect was significant** mainly **under CTR conditions**, while it was **largely reduced under** heat and feed **stress conditions**. Moreover, a **significant impact of feeding or temperature conditions** on microbiota richness and diversity was observed **in all lines**.

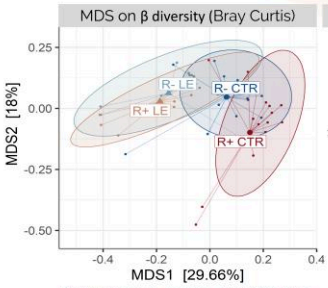


Permanova p-values < 0,001 for line, and condition.

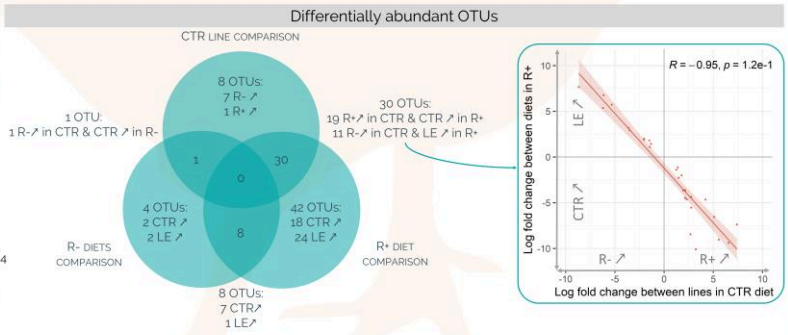
MICROBIOTA LINKS WITH FEED EFFICIENCY AND FEEDING CONDITIONS

Differential abundance (DA) analysis confirmed that lines differed almost exclusively under optimal feeding (CTR) and that the efficient line (R-) was much less impacted by the feed stress than the inefficient line (R+). Moreover, the **DA OTUs followed an interesting pattern: the more abundant OTUs in the R- line under CTR feed were often more abundant under LE diet in the R+ line**. This suggests that the microbiota of the R+ line feed with the LE diet, moved closer to the natural microbiota of the R- line, which is more diverse and probably underlines a better capacity to adapt to the diet change.

Among the taxonomies associated with these DA OTUs, 2 families were predominant: *Lachnospiraceae* and *Ruminococcaceae* with diverse genera such as *Faecalibacterium*, *Subdoligranulum* known to be short-chain fatty acid producers. However, based on their richness we cannot associate these genera to any of the conditions analysed. Among the other DA OTUs identified, those of the *Oscillospiraceae* family were preferentially more abundant in R- line and/or in LE diet, and on the contrary, the *Bifidobacteriaceae* and *Lactobacillaceae* families were more abundant in R+ line and/or in CTR diet.



Permanova p-values < 0,01 for line, and < 0,001 for diet.



At the functional level, we found same pattern of differences between lines under the CTR feed or between diets in the R+ line. The **R+ under CTR feed were characterised by** differential **functions involved in fatty acid biosynthesis, starch and sucrose, amino sugar and nucleotide sugar metabolisms**. In the **R- and the LE diet**, we found differential **functions involved in a** variety of metabolic pathways, including **carbohydrate metabolism** (in particular pyruvate, and glycolysis), **fatty acid biosynthesis and degradation,** and **protein metabolism with various amino acid metabolisms**.

Journées Scientifiques
Département Génétique animale 2022

Poster présenté lors des journées scientifiques du département de Génétique Animale 2022.

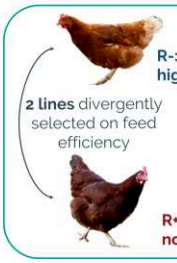
The richer the better: 16S metabarcoding analysis of gut microbiota of laying hens in relation to feed efficiency and adaptation to diet change

Bernard M.[1,2], Lecoeur A.[1], Coville J-L.[1], Bruneau N.[1], Jardet D.[1], Lagarrigue S.[3], Meynadier A.[4], Calenge F.[1], Pascal G.[4], Zerjal T.[1]

1: Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350, Jouy-en-Josas, France. 2: INRAE, SIGENAE, 78350, Jouy-en-Josas, France.
3: INRAE, INSTITUT AGRO, PEGASE UMR 1348, Saint-Gilles, France. 4: GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France

The gut microbiota is known to play an important role in energy harvest and is likely to affect feed efficiency. In the context of feed-food competition and where feed cost represents 70% of eggs production cost, studying the gut microbiota of laying hens allow to determine its role in feed efficiency and its sensitivity to the diet composition.

SCIENTIFIC QUESTIONS & EXPERIMENTAL DESIGN



How can microbiota influence the feed efficiency of laying hens ?

How are bacterial populations impacted by the diet composition ?

2 diets:
commercial diet, CTR (soybean + wheat based)
low-energy, LE (corn + sunflower based):
-15% metabolisable energy and
2.4 times more raw cellulose
-30% starch

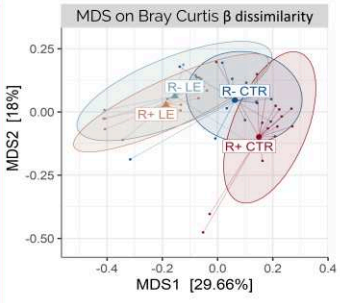
Material: 57 laying hens, 16Sv3V4 metabarcoding reads

Tools: FROGS[1], Phyloseq[2], DESeq2[3], PICRUSt2[4]

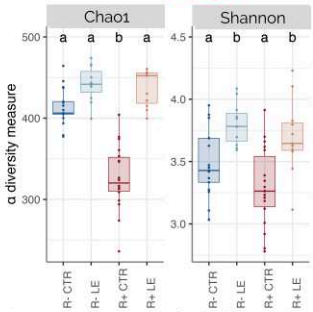
Bioinformatic output : Amplicon Sequence Variant (ASV) abundance table with taxonomic affiliation.

[1] Escudié, F. et al., Bioinformatics 2018; [2] McMurdie, P. J. & Holmes, S., PLOS One 2013; [3] Love, M. I. et al., Genome Biology 2014; [4] Douglas, G. M. et al., bioRxiv 2019

RESULTS



Permanova p-values < 0.01 for line, and < 0.001 for diet.



Anova p-values < 0.03 for the line, = 0 for the diet, and for the line x diet in Chao1.

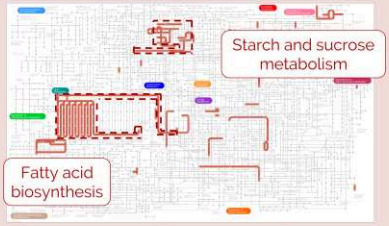
Alpha and beta diversity analyses, as well as differential ASV abundance analysis revealed :

- a **line effect** mostly observed **in the CTR diet:** difference in richness and structure between lines.
- a **diet effect** on the microbiota composition and structure is observed **in both lines** with an increase of diversity but with a **minor effect on the efficient line** R-.

Interestingly, taxonomies associated with **differentially abundant ASV and inferred functions** were shared between the R- efficient line fed the CTR diet and the LE diet groups. This **suggests** that **common** microbiota **mechanisms between feed efficiency and adaptation to non-traditional diet** exist.

R+ & CTR

- a greater abundance of *Actinobacteriota*, with in particular more abundant ***Bifidobacterium*** and ***Olsenella*** *ASVs, which* ***degrade sereval nature of carbohydrates, in particular starch.***

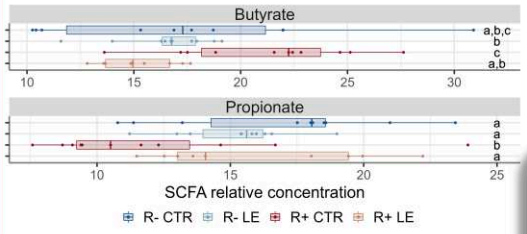


Hypothesis : The lower richness observed could illustrate an opportunistic state linked to the high amount starch reaching the caecum.

R- / LE

- a greater abundance of *Bacteroidota*, with in particular more abundant ***Alistipes***, ***Anaerosporobacter*** **or Bacteroides** ASVs, **known to degrade various complexe polysacharrides.**

Hypothesis : The higher and more diverse type of substrats, as undigested fibres, in the caecum is expected to increase the microbiota richness.



- The **greater** concentration of the **butyrate** in the **R+** line fed the **CTR** diet compared to the LE diet, **could be explained by more abundant** ***Firmicutes*** bacteria such as *Feacalibacterium* or *Subdoligranulum*.
- The **greater** concentration of the **propionate in the efficient** R- line or the LE diet, **corroborates** the functional inference and **the observed increase of glucose metabolism in the liver** of this line.

CONCLUSION

Our results are consistent with the idea that a **richer and more diverse microbiota** contributes to **improve animal feed efficiency** but this association is highly feed dependent. In optimal context, we hypothesise that efficient birds **optimise nutrients absorption**, through fibre degradation by fibrolytic bacteria that produce **propionate** and influence the host metabolism.

13th International Gut Microbiology Symposium 2023

Poster présenté lors du 13e symposium international de microbiologie intestinale, 2023.

A richer gut microbiota is related to better feed efficiency and diet adaptability in laying hens

M. Bernard[1], A. Lecoeur[1], J.L. Coville[1], N. Bruneau[1], D. Jardet[1], S. Lagarrigue[2], F. Calenge[1], G. Pascal[3] and T. Zerjal[1]
[1]Université Paris-Saclay, INRAE, AgroParisTech, GABI, Domaine de Vilvert, 78350 Jouy-en-Josas, France, [2]INRAE, Institut Agro, UMR PEGASE, 16, le clos, 35590 Saint-Gilles, France, [3]GenPhySE, Université de Toulouse, INRAE, ENVT, 24, chemin de Borde-Rouge, Auzeville Tolosane, 31326 Castanet Tolosan, France; maria.bernard@inrae.fr

In the egg industry, feed cost represents the majority of total production costs and breeding efforts are ongoing to improve feed efficiency of laying hens. The gut microbiota is known to play an important role in energy harvest and is likely to affect feed efficiency. In this study, we analysed the composition of caecal microbiota of 31 week old hens by 16S metabarcoding sequencing to characterise its composition, interactions with the host and influence on phenotypes of interest. As an animal model, we used hens of the R+ and R- lines divergently selected for high (low feed efficiency) and low (high feed efficiency) residual feed intake values respectively, that were fed either a commercial wheat-soybean diet (CTR) or a low-energy corn-sunflower diet (LE). Our results show a significant line effect on the microbiota richness and composition with the CTR diet, whereas with the LE diet, the microbiota was primarily affected by the diet change. A line × diet interaction was observed: the high efficient R- line presented a greater microbial richness and a reduced impact of the diet change compared to the low efficient R+ line. Interestingly, common taxonomies and/or predicted functions were highlighted between R+ and CTR diet, and between R- and LE diet, which could suggest that common microbiota mechanisms between feed efficiency and adaptation to nontraditional feedstuffs exist. OTUs of *Actinobacteriota* were more abundant in R+ birds and in the CTR diet, whereas OTUs of *Bacteroidota* were preferentially abundant in R- birds and/or LE diet. At functional level, carbohydrates and fatty acid metabolisms on the one hand, and short-chain fatty acids and amino acids metabolisms, on the other hand, were enriched in birds with low or high feed efficiency, respectively. These results provide insight into the role of the microbiota in laying hen feed efficiency and the impact of diet composition on microbiota.

Résumé de la présentation orale faite lors de l'EAAP 2023.

scientific reports

Check for updates

OPEN

Relationship between feed efficiency and gut microbiota in laying chickens under contrasting feeding conditions

Maria Bernard[1,2]✉, Alexandre Lecoeur[1], Jean-Luc Coville[1], Nicolas Bruneau[1], Deborah Jardet[1], Sandrine Lagarrigue[3], Annabelle Meynadier[4], Fanny Calenge[1], Géraldine Pascal[4] & Tatiana Zerjal[1]✉

The gut microbiota is known to play an important role in energy harvest and is likely to affect feed efficiency. In this study, we used 16S metabarcoding sequencing to analyse the caecal microbiota of laying hens from feed-efficient and non-efficient lines obtained by divergent selection for residual feed intake. The two lines were fed either a commercial wheat-soybean based diet (CTR) or a low-energy, high-fibre corn-sunflower diet (LE). The analysis revealed a significant line x diet interaction, highlighting distinct differences in microbial community composition between the two lines when hens were fed the CTR diet, and more muted differences when hens were fed the LE diet. Our results are consistent with the hypothesis that a richer and more diverse microbiota may play a role in enhancing feed efficiency, albeit in a diet-dependent manner. The taxonomic differences observed in the microbial composition seem to correlate with alterations in starch and fibre digestion as well as in the production of short-chain fatty acids. As a result, we hypothesise that efficient hens are able to optimise nutrient absorption through the activity of fibrolytic bacteria such as *Alistipes* or *Anaerosporobacter*, which, via their production of propionate, influence various aspects of host metabolism.

Chicken eggs are consumed worldwide and remain one of the most cost-effective animal-based sources of nutrition in the human diet. According to the OECD-FAO, global egg production has doubled since 1990 and is projected to increase by another 15% from 2021 to 2031[1]. In the chicken industry, feeding costs account for up to 70% of the total production cost[2], making feed efficiency a crucial phenotype for reducing production expenses and minimising feed waste and pollutants.

Feed efficiency measures the ability of an animal to make the most of its feed ration for maintenance and production, and in poultry, it is commonly evaluated using two indices. The feed conversion ratio (FCR) in layers measures the amount of feed required to produce one kilogram of eggs, while the residual feed intake (RFI) is the difference between an animal's observed feed intake (FI) and its predicted FI, which is statistically estimated based on maintenance and production requirements. RFI is considered a better proxy of feed efficiency than FCR as its estimation takes into account variations in maintenance energy expenditure[3].

Feed efficiency is a complex trait that is partly shaped by genetics[4], and selective efforts in chickens have targeted either modifications to the metabolism of birds[5] or their digestive capacity[6]. Environmental factors such as rearing conditions and diet composition also have an influence[7]. This is particularly significant given the globalisation of poultry production, which means that animals are exposed to a wide range of environments and feeding regimes. Recent research has highlighted the role of the gut microbiota in variations in digestive ability, and there is growing evidence that it can contribute to feed efficiency.

[1]INRAE, AgroParisTech, GABI, Université Paris-Saclay, 78350 Jouy-en-Josas, France. [2]INRAE, SIGENAE, 78350 Jouy-en-Josas, France. [3]INRAE, INSTITUT AGRO, PEGASE UMR 1348, Saint-Gilles, France. [4]GenPhySE, Université de Toulouse, INRAE, ENVT, 31326 Castanet-Tolosan, France. ✉email: maria.bernard@inrae.fr; tatiana.zerjal@inrae.fr

Indeed, the gut microbiota is known to be involved in a number of important processes, including immunity, behaviour, and digestion of feed for energy harvesting by the host[8,9]. The microbial communities present in the different compartments of the intestine enable the degradation of feed that is not digested by host enzymes[8]. Bacterial richness and diversity have been found to be particularly high in the caecum, where transit time is the longest compared to other intestinal segments[10,11] and major nutrients influencing the host metabolism[7,8,12,13]—such as vitamins B and K and short-chain fatty acids (SCFA)—are produced. The gut ecosystem evolves throughout the host's life and is particularly sensitive to various factors such as the genetics of the host, the rearing conditions, and feed formula[12,14].

Many studies have investigated the association between microbiota composition and feed efficiency in chickens[10,11,15–19], but the findings have been inconsistent. For example, Siegerstetter et al.[17] found reduced richness and diversity in the caecal microbiota of efficient broilers (based on RFI), whereas Stanley et al.[16] reported a richer microbiota in efficient broilers (based on FCR) but with similar diversity. Conversely, Yan et al.[11] found no difference in microbial diversity between poorly and highly efficient layers (based on RFI). These disparate findings may be explained by the extensive physiological and genetic differences between layers and broilers[12,20,21]. Information from the literature focuses mainly on broiler chickens and very little on layers[12,21], making it difficult to compare microbiota studies. The microbiota is also highly sensitive to environmental changes, so much so that even studies with multiple replicates under similar experimental conditions have not always achieved concordance among trials. Indeed, significant differences in richness and diversity were observed in one of the three trials of Stanley et al.[16], and in one of two geographical locations of Siegerstetter et al.[17]. Additionally, the lack of agreement among studies may also result from the limited ability to separate animals from standard populations based on their efficiency, as even the extremes of the feed efficiency distribution differ by relatively little[11,16,17].

In this study, we aimed to analyse the microbiota of adult laying hens to explore its contribution to chicken feed efficiency and investigate its sensitivity to feed changes. Specifically, we used a 16S rRNA gene metabarcoding approach to characterise the caecal microbiota of two RFI-divergent laying lines (R+ and R−) that are the result of more than 40 years of selection[22]. The advantage of using these lines is that, despite sharing a similar genetic background, their RFI values differ by more than five phenotypic standard deviations[3]. The two lines have comparable egg production and growth rates. However, thanks to this long-running programme of selection, they differ considerably in feed intake (56% higher in R+ than in R− hens) and other RFI-related traits such as fat deposition (higher in R− chickens) and diet-induced thermogenesis (higher in R+ chickens), indicating strong differences in energy metabolism between lines[3,23–25]. Hens were fed two diets, one representing the classical layer commercial diet (control, CTR) and the other a fibre-rich but energy-depleted diet (low-energy, LE) that might be found in countries with difficulties in accessing certain resources. This experimental design allowed us to (i) investigate the link between feed efficiency and the caecal microbiota, and (ii) explore the stability of this potential link under different feeding conditions.

Results

Line and diet effects on egg production and efficiency-related traits

Feed intake, feed efficiency indices (RFI and FCR), growth, and egg production traits were recorded for hens of the high efficient (R−) and the non-efficient (R+) lines fed either an optimal diet (CTR) or a low-energy, high-fibre diet (LE) as outlined in Table 1. Notably, significant differences in feed intake (FI) and efficiency indices were observed between the lines across both dietary conditions. Specifically, under the LE diet, the energy intake decreased by a 9% for the R+ and by a 15% for the R− lines compared to the CTR diet, despite no significant difference in FI between the two diets. LE-fed hens presented also reduced egg weights and yolk weights. However, no significant line or diet effects were observed for the other traits measured.

					ANOVA *P*-values	
Measures	**R− CTR**	**R+ CTR**	**R− LE**	**R+ LE**	**Line**	**Diet**
Feed Intake, FI (kg/28 d)	2.6 ± 0.1	4.0 ± 0.1	2.7 ± 0.1	4.1 ± 0.1	< 0.001*	0.390
Energy Intake (MJ/28 d)	29.4 ± 1.1	45.2 ± 1.1	25.2 ± 1.3	41.0 ± 1.3	< 0.001*	0.005*
Residual Feed Intake, RFI (kg/28 d)	− 0.6 ± 0.1	0.8 ± 0.1	− 0.3 ± 0.1	1.1 ± 0.1	< 0.001*	0.001*
Feed Conversion Rate, FCR (kg /28 d)	2.3 ± 0.1	3.6 ± 0.1	2.6 ± 0.1	3.8 ± 0.1	< 0.001*	0.039*
Egg laying rate (%)	86.6 ± 1.8	86.6 ± 1.8	84.9 ± 2.2	84.9 ± 2.2	1.000	0.475
Egg number	59.9 ± 1.9	59.9 ± 1.9	60.3 ± 2.2	60.3 ± 2.3	0.991	0.862
Egg weight (g)	48.6 ± 0.5	47.7 ± 0.5	47.2 ± 0.6	46.3 ± 0.6	0.176	0.050*
Egg mass (kg/28 d)	1.1 ± 0.0	1.1 ± 0.0	1.1 ± 0.0	1.1 ± 0.0	0.253	0.439
Yolk weight (g)	13.7 ± 0.2	13.4 ± 0.2	13.3 ± 0.2	13.0 ± 0.2	0.058	0.026*
Yolk proportion (%)	27.6 ± 0.3	27.8 ± 0.3	27.4 ± 0.3	27.6 ± 0.3	0.054	0.667
Body weight (kg)	2.0 ± 0.1	2.1 ± 0.1	1.9 ± 0.1	2.0 ± 0.1	0.058	0.071

Table 1. Least square means (± standard error) and significance estimates for line and diet effects on efficiency and production measures. R− and R+ are the efficient and non-efficient lines, respectively, and CTR and LE are the control and low-energy diets, respectively. Significant effects (*P*-value ≤ 0.05) are highlighted by an asterisk. The line x diet interaction was not significant and was thus removed from the model.

Microbiota composition, richness and diversity analyses

The 16S metabarcoding sequencing generated 3,077,915 paired-end reads, with between 25,783 and 78,980 read pairs per sample. From these sequences, we obtained 609 clusters as amplicon sequence variants (ASV) and eight of these were removed because of their weak prevalence. After processing, we obtained 601 ASVs, and each sample contained an average of 24,505 sequences. At the species level, 79.4% of the ASVs were affiliated to "unknown species" and 9.8% to multiple species. At the genus level, 39.4% were affiliated to "unknown genus" and 3.0% had ambiguous taxonomy. Considering the uninformative nature of most of the species-level affiliations, we used genus affiliations in all subsequent taxonomic analyses.

The two dominant phyla were *Bacteroidota*, and *Firmicutes* (Table 2). Within *Bacteroidota*, the most abundant families were *Bacteroidaceae* (36.6%), *Barnesiellaceae* (9.1%), and *Rikenellaceae* (4.7%), while the most abundant families of *Firmicutes* were *Ruminococcaceae* (23%) and *Lachnospiraceae* (7.7%). Proportions of *Firmicutes* were rather similar among the different line x diet groups. Instead, the relative abundance of *Bacteroidota* tended to be lower in the R+ line than in R− under both diets.

The *Actinobacteriota* were significantly influenced by line, being more abundant in the R+ line compared with the R− line, and by diet, being more abundant under the CTR diet compared with the LE diet. Finally, the *Proteobacterota* were more abundant in hens fed the CTR diet compared to those fed the LE diet.

The richness and diversity analyses revealed a significant line and diet effect on the microbiota (Fig. 1). Interestingly, the diet effect was milder in the R− line compared to the R+ line. More specifically, the change from the CTR to the LE diet induced a significant increase in richness in both lines but the range of change was not the same between lines, + 8% for the R− line and +35% for the R+ line. This explained the significant line x diet interaction observed (P-value < 0.001). The Shannon index values were also affected by line (P-value = 0.034)

Phylum	Nb of ASV	Mean ± SE	R− CTR	R+ CTR	R− LE	R+ LE	ANOVA P-values	
							Line	Diet
Bacteroidota	84	50.4 ± 0.2	53.1 ± 2.8	46.4 ± 2.7	54.9 ± 3.2	48.2 ± 3.2	0.052	0.613
Firmicutes	500	45.1 ± 0.2	42.8 ± 2.4	46.7 ± 2.3	43.5 ± 2.8	47.4 ± 2.8	0.186	0.828
Actinobacteriota	9	1.9 ± 0.4	1.8 ± 0.8	3.9 ± 0.7	0.0 ± 0.9	1.4 ± 0.9	0.027*	0.010*
Campylobacterota	2	1.4 ± 0.2	0.8 ± 0.4	1.4 ± 0.4	1.6 ± 0.5	2.2 ± 0.5	0.246	0.110
Proteobacteria	3	0.8 ± 0.1	1.1 ± 0.2	1.2 ± 0.2	0.3 ± 0.2	0.4 ± 0.2	0.549	0.000*
Desulfobacterota	3	0.4 ± 0.0	0.4 ± 0.0	0.4 ± 0.0	0.5 ± 0.1	0.5 ± 0.1	0.987	0.152

Table 2. Phylum relative abundance (± standard error) and significance estimates for line and diet effects. This table indicates the total number of ASVs per phylum, the global mean abundance, the least square mean abundance in each line x diet group ± standard error (SE), and the P-values for the effect of line (R− vs. R+) or diet (CTR for control vs. LE for low-energy) on phylum relative abundance. Significant effects (P-value ≤ 0.05) are highlighted by an asterisk. Line x diet interaction was not significant and was thus removed from the model.

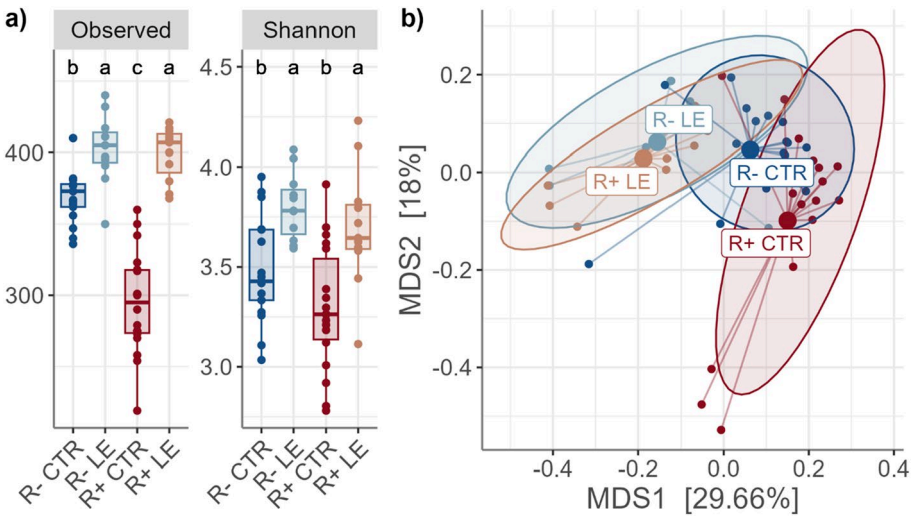


Figure 1. Alpha diversity distribution and beta diversity ordination plot: (**a**) Alpha diversity indices: observed richness and Shannon. In each panel, different letters at the top indicate significant pairwise differences (P-value ≤ 0.05) between line x diet groups. (**b**) Multidimensionnal scaling (MDS) of Bray Curtis dissimilarity between line x diet groups. Colours represent the four line x diet groups: red = R+ line fed the CTR diet; dark blue = R− line fed the CTR diet; orange = R+ line fed the LE diet; light blue = R− line fed the LE diet.

and diet (*P*-value < 0.001) (Fig. 1a). This trend was also evident in the multidimensional scaling (MDS) analysis of beta diversity (Fig. 1b), where distinct clustering patterns were observed. Specifically, the R− LE and R+ LE groups were clearly separated from the R+ CTR group, while the R− CTR group occupied an intermediary position. Statistical analyses validated these findings, confirming significant effects for both line (*P*-value = 0.007) and diet (*P*-value < 0.001).

Identification of taxonomic and functional differences in microbiota

The differential abundance analysis allowed us to identify the ASVs contributing to the differences between line x diet groups. The four comparison groups consisted of: (i) R+ versus R− hens fed the CTR diet, (ii) R+ versus R− hens fed the LE diet, (iii) R + hens fed the LE diet versus the R+ hens fed the CTR diet, and (iv) R− hens fed the LE diet versus the R− hens fed the CTR diet. Supplementary Table S1 provides the full results for the 94 Differential Abundant (DA) ASV in at least one of the four comparisons.

When examining the diet effect (LE versus CTR), we observed a substantial discrepancy in the number of DA ASVs between the R+ and R− lines. Specifically, the R+ line presented the highest number of DA ASVs (80), whereas the effect was smaller in the R− line (13). On the other hand, when analysing the line effect (R+ versus R−), we observed the highest number of DA ASVs under the CTR diet (39), whereas the effect was minimal under the LE diet (1 DA ASVs). These results reinforced the pattern observed previously highlighting the impact of the diet in both lines, albeit to a lesser extent in the R- line, and a line effect particularly evident in the CTR diet.

By investigating ASVs affected by both line and diet, we could identify ASVs that were shared between groups and others that were specific to one group. Interestingly, we identified 30 ASVs that were DA both between the R+ CTR versus R- CTR, and between the R+ LE versus R+ CTR (Fig. 2a). Among these, 11 ASVs presented higher abundance in R- CTR compared with R+ CTR and in R+ LE compared with R+ CTR. This displayed that ASVs that were more abundant in the R- line compared with the R+ line under standard conditions (CTR) become more abundant in the R+ hens when fed the LE diet (Fig. 2b). Conversely, the remaining 19 DA ASV were, specific to the R+ CTR group.

Distribution of bacterial taxonomies among line and diet comparisons

In terms of taxonomic distribution, the vast majority of DA ASVs belonged to the phylum of *Firmicutes* (77 ASVs), with a significant portion (93%) assigned to the order *Clostridia*. The remaining 17 DA ASVs were distributed across *Bacteroidota* (7%), *Actinobacteriota* (4%), *Proteobacteria* (3%), *Desulfobacterota* (2%), and *Campylobacterota* (1%). Taxonomic assignments are presented in Fig. 3 and detailed in Supplementary Table S1.

Many taxonomic groups exhibiteda pattern similar to that described for the DA ASVs, highlighting the distinctiveness of the R+ CTR group from the other groups. Some taxa were associated with DA ASVs specifically associated with the R+ CTR group, and conversely, others showed higher abundace in the R- CTR and in R+ LE groups. In particular, DA ASVs associated with the phylum *Actinobacteriota* were more abundant in the R+ CTR group and three of these, from the *Bifidobacterium* and *Olsenella* genera, presented mean relative abundances between 0.5 and 3%. Conversely, the DA ASVs associated with phylum *Bacteroidota* were generally more abundant in the R- CTR group (e.g., genus *Alistipes*) or in the R+ LE group (e.g., genus *Bacteroides*). Notably, the change in abundance of DA ASV affiliated to the latter was remarkable and passed from a relative abundance of 4.3% in the R+ CTR to 18.2% in the R+ LE group. Other taxa were associated to ASVs that were differential in several groups. For example, the genus *Faecalibacterium* was associated with three DA ASVs that were more abundant in the R- CTR group and five DA ASVs that were more abundant in the R+ CTR group.

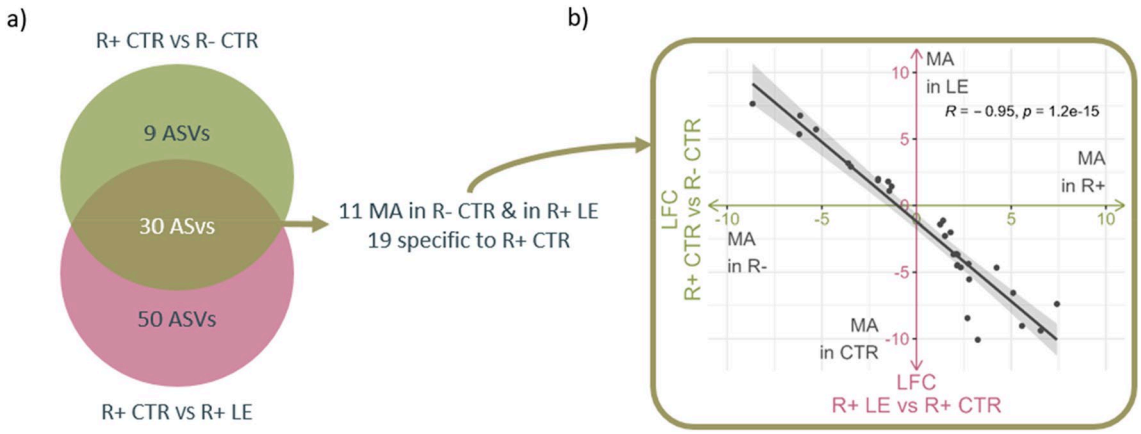


Figure 2. Venn diagram and log fold change of differentially abundant (DA) ASVs. (**a**) Venn diagram depicting unique and shared DA ASV between those identified from the R+ CTR versus R− CTR comparison and those identified from the R+ LE versus R+ CTR comparison; (**b**) Log-fold change (LFC) of shared DA ASVs, represented with black dots, between the two comparisons. Eleven ASVs that were more abundant (MA) in the R− fed CTR diet were also more abundant in the R+ fed the LE (top left corner of the plot). Conversely, 19 ASVs were specific to the R+ CTR group (bottom right corner of the plot).

					CTR line comparison		R- diet comparison		R+ diet comparison	
Phylum	**Class**	**Order**	**Family**	**Genus**	**R-**	**R+**	**CTR**	**LE**	**CTR**	**LE**
Actinobacteriota	Actinobacteria	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium	0	3	0	0	3	0
	Coriobacteriia	Coriobacteriales	Atopobiaceae	Olsenella	0	1	0	0	1	0
Bacteroidota	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides	0	0	0	0	0	2
			Barnesiellaceae	Barnesiella	0	0	0	0	1	0
				unknown genus	1	0	0	0	0	1
			Marinifilaceae	Odoribacter	0	0	0	0	0	1
			Rikenellaceae	Alistipes	2	0	0	0	0	0
Campylobacterota	Campylobacteria	Campylobacterales	Campylobacteraceae	Campylobacter	0	0	0	0	1	0
Desulfobacterota	Desulfovibrionia	Desulfovibrionales	Desulfovibrionaceae	Desulfovibrio	0	0	0	0	1	0
				unknown genus	0	0	0	0	0	1
Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus	0	0	2	0	1	0
				Ligilactobacillus	0	0	0	0	1	0
				Limosilactobacillus	0	1	1	0	2	0
	Clostridia	Clostridia UCG-014	unknown family	unknown genus	3	1	0	0	3	7
		Clostridia vadinBB60 group	unknown family	unknown genus	2	0	0	0	0	3
		Lachnospirales	Lachnospiraceae	[Ruminococcus] torques group	0	1	0	0	3	0
				Anaerosporobacter	1	0	0	0	0	1
				Anaerostipes	0	0	0	0	1	0
				Eisenbergiella	1	0	1	1	1	2
				Fusicatenibacter	0	2	0	0	2	0
				GCA-900066575	0	0	0	0	0	1
				Lachnospiraceae FE2018 group	0	0	1	0	1	0
				Multi-affiliation	1	0	0	0	0	2
				Sellimonas	0	1	0	0	1	0
				unknown genus	1	1	1	0	2	4
		Oscillospirales	[Eubacterium] coprostanoligenes group	unknown genus	0	0	1	0	0	0
			Butyricicoccaceae	Butyricicoccus	0	0	0	0	1	0
			Oscillospiraceae	Colidextribacter	0	0	0	0	0	1
				Multi-affiliation	1	0	0	0	0	1
				NK4A214 group	0	0	0	0	0	0
				Oscillibacter	0	0	0	0	1	0
				UCG-005	0	0	0	0	0	1
				unknown genus	1	0	0	1	0	2
			Ruminococcaceae	Caproiciproducens	1	1	0	0	1	0
				DTU089	1	0	0	0	0	0
				Faecalibacterium	3	5	1	0	8	2
				Fournierella	0	0	0	0	0	1
				Negativibacillus	0	0	0	1	0	1
				Ruminococcus	0	0	0	0	1	1
				Subdoligranulum	0	2	1	0	4	1
		Peptostreptococcales-Tissierellales	Peptostreptococcaceae	Romboutsia	0	0	1	0	1	0
Proteobacteria	Gammaproteobacteria	Burkholderiales	Sutterellaceae	Parasutterella	0	0	0	0	2	0
		Enterobacterales	Enterobacteriaceae	Escherichia-Shigella	0	1	0	0	0	0

Figure 3. Heatmap illustrating the number of ASVs significantly more abundant in one group within each comparison. The CTR line comparison displays ASVs identified as differentially abundant in the R+ CTR group (depicted in red) versus the R− CTR group (depicted in blue). The R− and R+ diet comparisons reports ASVs differentially abundant in the R− LE group (depicted in purple) versus the R− CTR group (depicted in green), or in the R+ LE (depicted in purple) versus R+ CTR (depicted in green).

The comparison between the CTR diet and the LE diet also revealed abundance changes in several other genera belonging to the families *Lactobacillaceae*, *Lachnospiraceae*, or *Ruminococcaceae*.

Comparison of inferred microbial functions

KEGG functional inference allowed to identify 4,403 functions. Differential function analysis revealed that the greatest function shift occurred between the R+ CTR and the R+ LE groups, presenting 841 DA functions. In contrast, the change between CTR to LE diets in the R- line was less dramatic, with 217 DA functions identified. The functional differences between R+ and R− lines were mainly observed under the CTR diet, with 329 DA functions identified. Conversely, under the LE diet, the functional profiles of the two lines showed a high degree of similarity, with only 43 DA functions detected (Supplementary Table S2).

The DA functions identified in the R+ CTR versus R- CTR comparison could be associated with a large number of pathways, some more abundant in R+ line and others in the R- line (Fig. 4, and supplementary Table S3). In particular, functions more abundant in the R- CTR microbiota were associated with various amino acid metabolisms and degradations, as well as butanoate and propanoate metabolisms. In contrast, the R+ CTR microbiota appeared to be enriched in the metabolism of various carbohydrates, notably starch and sucrose, and galactose.

The similarity between the R− CTR group and the R+ LE group, already observed at the level of DA ASVs, is also observed at functional level. Indeed, pathways that appeared to be more active in the R− CTR group were often also more active in the R+ LE group. On the other hand, pathways related to carbohydrates metabolisms, including starch and sucrose metabolism, which were over-represented in the R+ CTR group, were specific to this line and diet condition.

Line and diet effects on relative concentrations of microbial short chain fatty acids (SCFA)

The quantification of acetate, propionate, and butyrate revealed that acetate was the most abundant SCFA, accounting on average for 67% of the total SCFA concentration, followed by butyrate (18%) and propionate (15%) (Table 3 and supplementary Figure S1).

No significant differences were detected between groups for acetate. However, variations in propionate and butyrate concentrations were notable within the R+ CTR group. The relative concentration of propionate was lower in the R+ CTR group compared to the other groups. Conversely, the relative concentration of the butyrate was significantly higher in the R+ CTR group compared to R+ LE and R− LE groups. While it also tended to be higher than in the R- CTR group, the substantial variance observed in the latter group precludes definitive conclusions.

Correlation values between the relative concentrations of SCFAs and the relative abundances of families and genera are presented in supplementary Tables S4 and S5, and significant correlations are shown in Fig. 5.

A negative correlation of −0.6 (P-value < 0.001) was observed between propionate and butyrate, as depicted in Fig. 5, where several genera exhibit contrasting correlated values with these two SCFAs. For example, the *Subdoligranulum* and *Lactobacillus* were positively correlated with butyrate, while they were negatively correlated with propionate, and an opposite correlation trend was observed for the genus *Tuzzerella*. Other genera were correlated with only one SCFA. For example, the genus *Bacteroides* was positively correlated with propionate, while *Fusicatenibacter* was negatively correlated.

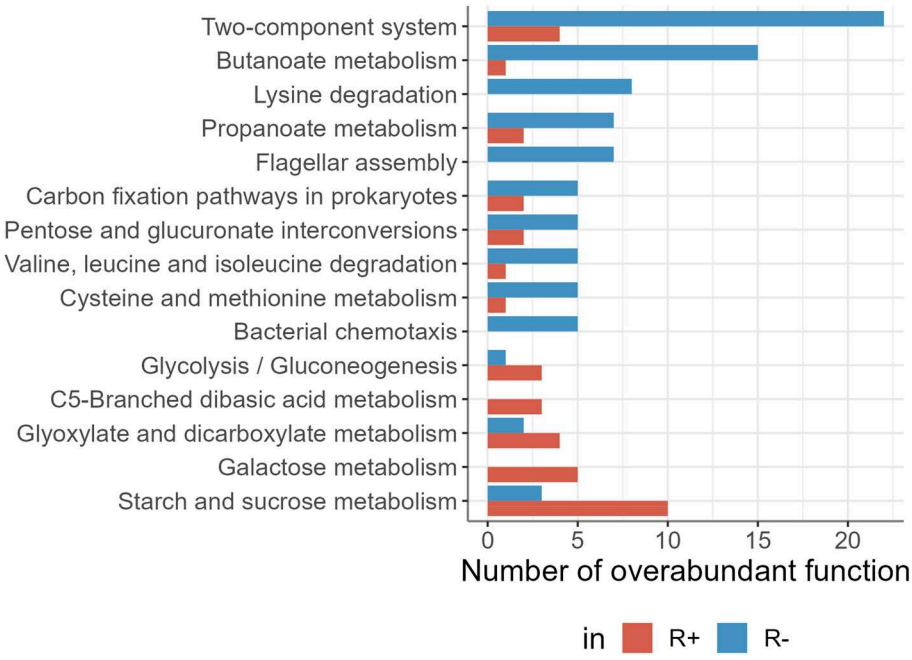


Figure 4. Pathway enriched in differentially abundant functions. Representation of the top 15 represented pathways associated with at least three DA functions. The R+ CTR group is depicted in red and in the R− CTR group in blue.

SCFA	R– CTR	R– LE	R+ CTR	R+ LE	ANOVA *P*-values		
					Line	Diet	Line x Diet
Acetate (%)	64.9 ± 4.2[a]	68.2 ± 3.5[a]	66.7 ± 4.0[a]	68.7 ± 3.5[a]	0.500	0.053	0.796
Propionate (%)	16.7 ± 4.0[a]	15.1 ± 2.1[a]	12.3 ± 4.3[b]	16.0 ± 3.8[a]	0.142	0.397	0.020*
Butyrate (%)	18.4 ± 6.9[a,b]	16.7 ± 2.2[b]	21.0 ± 4.4[a]	15.2 ± 1.7[b]	0.325	0.010*	0.053

Table 3. Mean and standard deviation of SCFA relative concentration and significance of line and diet effects. ANOVA *P*-values for line effect, diet effect, and interaction were considered significant if ≤ 0.05 and were highlighted with an asterisk. For each SCFA, the relative abundances with different superscripts differed significantly (*P*-value ≤ 0.05).

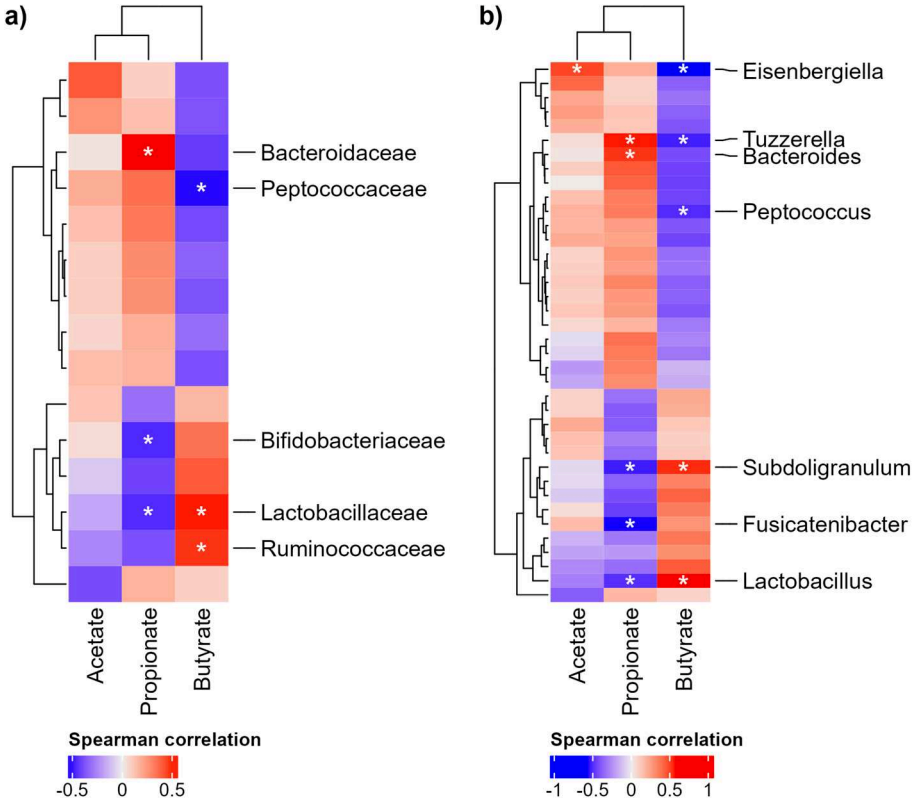


Figure 5. Correlation analysis between relative concentrations of SCFAs and relative abundances of bacterial taxa. (**a**) Correlations at the family level, (**b**) correlations at the genus level. Negative Spearman correlation values below −0.3 are depicted in blue, while positive values above 0.3 are shown in red. Significant correlations (adjusted *P*-values ≤ 0.05) are indicated with asterisks, along with the associated taxonomies.

Discussion

A better understanding of the role played by the caecal microbiota in chicken feed efficiency could help improve chickens' ability to extract nutrients from feed. This could have both economic and environmental benefits, including reduction in feed costs and the limitation of feed waste and pollutants.

In this study, we analysed adult laying hens from two experimental lines, one characterised by a low feed efficiency (R+) and the other by a high feed efficiency (R−)[22]. These lines were fed either a standard laying commercial diet (CTR) or a low-energy, high-fibre diet (LE). Previous heritability analyses in these lines have indicated a moderate heritability for RFI (between 0.33 and 0.27)[26], suggesting that this trait is partly under genetic control and partly controlled by other factors. One such factor could be the influence of the digestive microbiota, which is known to play a role in feed degradation and host metabolism[27]. To explore this aspect, we performed 16S metabarcoding sequencing to characterise the diversity of caecal microbiota in R+ and R− hens across both feeding conditions, and to identify the microbiota functions distinguishing these lines. As in other studies in adult chickens[28–30] the dominant phyla were *Firmicutes* (45%) and *Bacteroidota* (50%).

Feed driven changes in microbiota diversity in R+ and R− chicken lines

Our study corroborated previous findings highlighting the significant impact of diet composition on microbiota diversity and richness[21], in particular emphasizing the central role of fibre content[31]. The LE diet used in this

experiment, compared with the CTR diet, was characterised by a 2.1-fold increase in fibre content from various cereal sources, along with a 23% reduction in starch content, resulting in lower metabolisable energy and lower digestibility. Fibres are not digested by the host's digestive enzymes in the small intestine, but undergo partial degradation via fermentation by the caecal microbiota in the large intestine. Fibre composition is of major importance, as soluble fibres in particular, provide a wide range of substrates for fermentation reactions carried out by dedicated microbes, consequently influencing microbiota composition, diversity, and richness[32,33]. In our study, the increase in microbial richness and diversity in hens fed the LE diet is likely due to the wider variety of fibre types in this diet, as the physiochemical properties of different fibres, such as solubility, viscosity, and fermentability, vary considerably depending on their origin[31]. However, this greater variety of fibres affected the R+ and R− lines differently, as shown by the difference in microbial response between the two lines under the LE diet. The microbiota of the efficient R- line was less affected by the change in diet, probably due to its natural tendency for higher microbial richness and a more efficient use of fibre even under the regular CTR diet. This is in line with what reported for the human microbiota, where richer microbiota remained stable despite an increased fibre content in the diet[34].

In contrast, the microbiota of the R+ line was more affected by the LE diet; the LE formula appeared to drive a dynamic shift that brought the composition of the R+ microbiota closer to that of the R- line. It is highly probable that similar substrates reaching the caecum resulted in similar microbiota compositions between the R+ and R− lines under the LE diet, unlike the scenario observed under the CTR diet.

Interestingly, several ASVs that were more abundant in the R− CTR group were also more abundant in the R+ LE group. This was also observed at the functional level, with numerous functions being more prevalent in both the R− CTR and R+ LE groups. This pattern probably reflects the pressure exerted by the LE diet on the R+ microbiota to expand its functional range in order to degrade the wider variety of substrates offered by a high-fibre regime, activating functions already observed in the R− line under the CTR feeding condition.

Microbiota richness and diversity differences between efficient and non-efficient hens

A notable finding of our study concerns the genetic effect observed on microbiota richness and diversity, with diet acting as a modulating factor. Particularly, a significant genetic effect was observed under the CTR diet, whereas minimal microbial disparities between efficient and non-efficient hens existed under the LE diet.

The efficiency of digestion and energy harvesting from feed has been associated with the activity and composition of the gut microbiota[35,36]. In this context, the increased caecal microbiota richness observed in the R− line may reflect enhanced nutrient utilisation from feed, potentially contributing to the improved efficiency of R− hens, which are able to maintain production and growth despite a significant reduction in feed intake compared to R+ hens. On the other hand, the specificity of the R+ microbiota under the CTR diet may be related with the higher feed intake characteristic of this line (on average 56% more than R− line). Indeed, ingestion levels and eating patterns have been shown to shape microbiota composition[37]. One hypothesis could be that in the R+ line, excessive feed intake may lead to reduced starch digestibility. While starch is generally well digested by the chicken intestine, some studies have reported a negative correlation between starch digestibility and starch intake, particularly in wheat-based diets[38]. Moreover, wheat is also rich in arabinoxylan, a soluble fibre that can increase viscosity when present in excess and by consequence reduce nutrient digestion[31]. This could result in an overabundance of resistant starch reaching the caecum. Resistant starch is highly fermentable and, like soluble fibres, can modulate microbiota composition. The R+ microbial community might reflect a form of bacterial opportunism prompted by excessive quantities of resistant starch. Indeed, a study on pigs has shown that the composition of the microbiota can be modified in favour of starch metabolism, thereby delaying the fermentation of other soluble fibres like arabinoxylan and β-glucan[39].

Overall, the observed variations in richness and diversity within the microbiota of efficient and non-efficient hens seem to suggest the presence of two distinct microbial ecosystems. One appears specific to the non-efficient R+ line under the CTR diet, and may be shaped by starch fermentation. The other ecosystem, observed in the efficient R- line under the CTR diet as well as in hens fed the LE diet, is likely capable of degrading a variety of carbohydrates and utilizing a wider range of fibre types as substrates.

Increased abundance of bacterial clades related to residual starch fermentation in the non-efficient R+ line under the CTR diet

The R+ CTR group is characterised by an abundance of ASVs associated with *Bifidobacterium*. This specificity of *Bifidobacterium* for the R+ CTR group is substantial considering that of the five ASVs associated with this genus, four were not detected in the R- line or in any hen fed the LE diet. We were able to affiliate one of these ASVs to the species *Bifidobacterium pseudolongum* that was particularly abundant in the R+ CTR group (0.5%). Increase in *Bifidobacterium* content was observed often in relation with resistant starch fermentation[40,41]. The starch contained in cereal grains is the main energy source in poultry diets. In the two diets used in this study, starch was 23% lower in the LE diet than CTR diet. Due to the larger feed intake of R+ hens, it is probable that some undigested starch reaches the caecum, leading to starch fermentation. In pigs, this has been correlated with a lower feed efficiency and with an increased population of *Bifidobacterium*[42], which is in line with our results. The preferential growth of *Bifidobacterium* over other genera during starch fermentation likely occurs because certain strains of *Bifidobacterium* possess the necessary enzymatic activity to utilise starch specifically in the distal gut[43]. Interestingly, we detected increased activity of pathways related to starch and sucrose metabolism in the R+ line under the CTR diet, which supports our hypothesis of differences in the microbial utilisation of starch between the two lines.

Wheat, which largely composes the CTR diet (600g/kg) is also rich in arabinoxylan, a soluble fibre well fermented by the gut microbiota. The fermentation by the microbiota of resistant starch and dietary fibres results

mainly in the production of SCFAs, and in the case of resistant starch and arabinoxylan, the fermentation is in favour of the butyrate production[39]. This could explain the observed higher proportion of butyrate produced by the R+ CTR microbiota compared with the other groups.

Butyrate and genus *Bifidobacterium* are generally considered to have beneficial effect on gut health, by improving gut barrier function and maintaining gut homeostasis through cross-feeding interactions with other bacteria[12,44]. Interestingly, the R+ line has been reported to have better immune functions compared to the R- line, presenting lower mortality rate following infection and increased antibody production[3].

In the R+ CTR group, there were several other ASVs affiliated to SCFA-producing genera, such as *Olsenella* and *Fusicatenibacter*, which are known acetate producers[45,46] and have previously been associated with resistant starch consumption[47,48]. Numerous ASVs belonging to the genera *Faecalibacterium*, *Subdoligranulum*, and *Fusicatenibacter* (phylum *Firmicutes*) were also more abundant in the R+ CTR group. These genera are known butyrate producers[49,50] and in our study, the abundance of *Subdoligranulum* was found to be positively correlated with the relative concentration of butyrate. Previous research has identified cross-feeding interactions between *Bifidobacterium* and *Faecalibacterium*[51], which could explain the abundance of both genera in the R+ CTR group. In another study, *Faecalibacterium* was associated with a low feed efficiency in laying hens[11], which is consistent with our results.

Increased abundance of propionate-producing bacteria in the efficient R- line and under LE feeding conditions

Interpreting the patterns of ASVs abundance in the R- line or in the LE diet groups is a complex task, as the majority of the DA ASVs were associated with unknown genera or had multiple affiliations. Nevertheless, among the ASVs that could be associated with known genera, we identified *Anaerosporobacter* as being more abundant in the R- CTR group as well as in the R+ LE group. This genus has been described as cellulose degrader[52], which may explain its increase in the R+ LE group. Furthermore, it has recently been reported to be negatively correlated with RFI and to contribute to improved feed efficiency in adult layers[53], in lines with our observation of increase in the R- CTR group. We also identified *Negativibacillus* as being more abundant in both lines in LE diet; this genus being implicated in the digestion of complex carbohydrates[54].

It also emerged that the large majority of DA ASVs associated with phylum *Bacteroidota* were more abundant in the R− CTR group and/or in the R+ LE group. The role of these bacteria as propionate producers is corroborated by the presence in this phylum of enzymes involved in the succinate pathway, which is the main route for propionate production from ingested carbohydrates[55]. For example, two DA ASVs more abundant in the R- CTR than in the R+ CTR group were associated with genus *Alistipes*. Previous research has linked this genus with improved digestibility[18] and feed efficiency in broiler chickens and pigs[56–58]. Our results in laying hens are consistent with these earlier findings and support the hypothesis that *Alistipes* may play a role in improving feed efficiency. The *Alistipes* genus is known to degrade cellulose[59], yet it was not detected as more abundant in the LE diet groups, which is surprising. This observed lack of increase in *Alistipes* abundance under the LE diet may indicate a competitive dynamic among cellulolytic bacteria for the same substrate that will favour the more abundant ones[32]. Indeed, *Bacteroides*, also known to degrade cellulose and to produce propionate[60], were more abundant than *Alistipes* in the CTR groups and they increased significantly in the R+ LE group. They were also positively correlated with the propionate relative concentration, which was higher in the LE groups.

The potential role of propionate-producing bacteria in enhancing feed efficiency is further supported by the functional analysis, which revealed an activation of the propanoate metabolism pathway in the R− CTR group. This finding is corroborated by the higher relative concentration of propionate in this group compared to the R+ CTR group. This evidence supports the hypothesis that propionate could represent an extra energy source for the host, potentially contributing to improve feed efficiency. Notably, propionate in the liver acts as a precursor for hepatic gluconeogenesis[61], a process found to be enhanced in the R- line compared to the R+ line, as revealed by the overexpression of key gluconeogenesis genes (Jehl et al. in preparation). It is highly plausible that the additional energy derived from microbial propionate supports the nutritional needs of the R− line, helping to compensate for its greatly reduced feed intake, allowing it to maintain metabolic energy homeostasis.

Conclusion

This study provides a clear demonstration of the usefulness of the R+ and R− experimental chicken lines, which diverge dramatically with respect to RFI, as models for investigating the role of the gut microbiota in shaping feed efficiency under different feeding conditions. Our results are consistent with the idea that a richer and more diverse microbiota is associated with improvements in hen feed efficiency. It is likely that the long-term regime of intensive selection on high or low RFI values has contributed, either directly or indirectly, to the observed differentiation through the establishment of line-specific gut ecosystems. However, this microbial differentiation between lines could only be detected under the optimal feeding conditions used for the RFI-divergent selection, indicating that the association between the microbiota and feed efficiency is highly feed-dependent. Additionally, we observed that certain taxa and/or inferred functions were shared between the R− CTR group and both LE diet groups, suggesting the existence of common microbiota mechanisms to optimise energy extraction from feed. One such mechanism may be an increase in the abundance of propionate-producing bacteria in order to provide extra energy to host metabolic functions. These results offer insights into how the feed efficiency of laying hens may be influenced by the gut microbiota and, in turn, how these microbial communities are affected by diet composition.

Methods

Ethics statement

The experiment was conducted at the experimental farm PEAT (INRAE, Val-de-Loire Center, Nouzilly) under licence number C37-175-1 for animal experimentation, in compliance with European Union Legislation, and was approved by the local ethics committee for animal experimentation (Val de Loire) and by the French Ministries of Higher Education and Scientific Research, and of Agriculture and Fisheries (n°2873-2015112512076871). The study followed the ARRIVE guidelines.

Animal information and sample collection

This study involved hens from two chicken lines that have been divergently selected for residual feed intake (RFI). RFI represents the deviation of the observed feed intake (FI) of an animal from its predicted feed intake (PFI) calculated based on maintenance and production requirements. These two experimental lines were derived from the same population of Rhode Island Red chickens, and have been developed by the French National Research Institute for Agriculture, Food, and the Environment (INRAE) since 1976 as described in Bordas et al.[22]. This divergent selection program has resulted in an efficient line, named R-, which eats less than estimated (RFI < 0), and an inefficient line, named R+, which eats more than estimated (RFI > 0) despite no differences in terms of growth or eggs production. From this cohort, the present study focused on 58 hens fed two distinct diets. All hens used in this experiment shared the same environment from hatching to sample collection at 31 weeks of age. All eggs were hatched the same day in the same hatchery and chicks were reared together in floor pens under standard rearing conditions until 17 weeks of age. Then, they were transferred into individual cages, with a light regimen set at 14 h of light per day and ad libitum feeding. Of the 58 hens used in this study, 18 hens per line were fed a commercial diet (control group, CTR) and 11 per line were fed a low-energy diet (low-energy group, LE) until they reached 31 weeks of age. At that point, the hens were fed, subjected to head electrical stunning, and immediately slaughtered by neck cut and bleeding. Caecal content was collected shortly after, and samples were snap-frozen in liquid nitrogen and conserved at – 80 °C until DNA extraction was performed.

The two diet formulas, detailed in Table 4, contained similar protein content but differed in energy content. The LE diet was 15% lower in metabolisable energy compared to the CTR diet (9.7 MJ/kg vs. 11.3 MJ/kg). Additionally the LE diet had 23% lower starch content (302 g vs. 393 g) but was 2.1 times higher in raw cellulose (63.2 g/kg vs. 29.6 g/kg). This was due to the reduction of wheat and soybean and to the increase of corn, sunflower, rapeseed, and oat. Metabolisable energy, protein and starch content were confirmed by chemical analyses, and cellulose, hemicellulose and lignin were estimated using Van Soest formulas[62].

Phenotypic trait collection

Individual traits related to feed efficiency, egg production, and growth were recorded for the 58 hens of the study. Egg number was recorded from 21 to 31 weeks of age and the egg laying rate was estimated as (number of eggs/recorded period)*100. Individual feed intake (FI) was recorded over a four-week period from 28 to 31 weeks of age. Energy intake was calculated by multiplying the FI (g/28 days) by the metabolisable energy of the diets (11.3 MJ/kg and 9.7 MJ/kg for the CTR and LE diets, respectively (Table 4)). Body weight was recorded at 28 (BW28) and 31 weeks (BW31) of age and the body weight gain estimated as BW31–BW28. Egg mass was estimated by adding up the weight of eggs laid over the 28 days recording period. RFI was calculated as the difference between the observed FI and the predicted FI (PFI) for the recorded period: (RFI = FI – PFI). PFI was estimated by a multiple regression equation calculated for all hens using three independent variables: average body weight (BW), BW gain, and egg mass produced over the recorded period[22]. The feed conversion ratio was estimated between 28 and 31 weeks of age as the ratio between the total feed intake and the egg weight produced over that period. Yolk weight was estimated from three eggs per hen collected at week 31 and yolk proportion estimated as (average yolk weight/average egg weight)*100.

Short-chain fatty acids (SCFAs) in caecal samples

The SCFA (acetate, propionate, and butyrate) content of 40 caecal samples (10 from R– CTR, 10 from R– LE, 11 from R+ CTR, and 9 from R+ LE) was determined by gas chromatography following the protocol previously described by Bedu-Ferrari et al.[63]. Between 100 and 250 mg of caecal content were diluted in two volumes of deionised water. Samples were homogenised and mixed for 2 h at 4 °C before centrifugation at 12,000 × *g* for 15 min at 4 °C. The supernatant was then collected and weighed, and 10% (vol/vol) of phosphotungstic acid saturated solution (Sigma-Aldrich) was added for protein precipitation overnight at 4 °C. As an internal standard, 10 µL of 2-ethylbutyrate (Sigma-Aldrich) were added to 40 µL of acidified supernatant, and the solution was analysed using a gas–liquid chromatograph (GC-FID Agilent 7890B). All samples were analysed in duplicate. Data were collected and peaks were integrated using Agilent OpenLab Chemstation software. Prior to modelling, the relative concentrations of the three SCFAs were computed as the SCFA concentration (in µmol/g) divided by the total concentration of the three main SCFAs (acetate, butyrate, and propionate).

Metabarcoding sequencing

DNA was extracted from frozen caecal content following the protocol described by Gordon et al.[64] at the @bridge platform (INRAE, Ile-de-France, Jouy-en-Josas). Subsequently, the V3–V4 hyper-variable regions of the 16S rRNA gene were amplified with two rounds of PCR following the protocol described in Lluch et al.[65] with the modified primers PCR1F_460 (ACGGRAGGCAGCAG) and PCR1R_460 (TACCAGGGTATCTAATCCT)[66]. Paired-end sequencing (2 × 250 pb) was performed on the Illumina MiSeq Platform.

	Diets	
	CTR	LE
Ingredients (g/kg)		
Wheat	599.0	100.0
Corn	50.0	332.5
Cereal co-products	0.0	80.0
Soybean meal	197.6	92.1
Sunflower meal	0.0	152.9
Rapeseed meal	0.0	50.0
Rapeseed seeds	20.0	0.0
Oat	0.0	55.7
Corn gluten meal	2.0	0.0
Alfalfa protein concentrate	5.0	0.0
Soybean oil	10.8	20.0
Minerals, vitamins and pigments	113.0	112.0
Lysine HCl	1.0	1.5
DL-Methionine	0.6	1.1
L-Threonine	0.0	0.2
Energy and nutrient contents		
Metabolisable energy (MJ/kg)	11.3	9.7
Protein (g/kg)	165.0	158.0
Ca (g/kg)	40.0	40.0
Available P (g/kg)	3.4	3.0
Lysine (%)	0.9	0.8
Methionine (%)	0.4	0.4
Starch (g/kg)	393.0	302.0
Cellulose*	29.6	63.2
Hemicellulose*	73.2	95.7
Lignin*	8.9	26.4

Table 4. Composition of CTR and LE diets. CTR correspond to the control diet, and LE to the low-energy, high-fibre diet. Asterisks indicated estimated values using the Van Soest formulas[62].

Sequence analysis

After removing one sample because of low numbers of raw reads (1108 raw pairs), the sequences of the remaining 57 samples were analysed with FROGS[67] (version 3.2.3) following up-to-date pipeline guidelines (http://frogs.toulouse.inra.fr/). The specific settings used for the tools were as follows: (i) for read pair assembling with Pear[68] (version 0.9.10), amplicon size range was set between 300 and 490 bp, (ii) for clustering, the *-d1-fastidious* options were used, (iii) after removing chimeras, we kept clusters with a minimum relative abundance of 0.005% (i.e., more than 96 sequences across all 57 samples), as suggested by Bokulich et al.[69] and present in at least 30% in at least one all line x diet group, and (iv) for taxonomic affiliation we used the 16S SILVA database (version 138.1)[70] with a minimal pintail score of 50. Functional analysis was also performed using FROGS (version 4.0.1), which relies on the PICRUSt2 suite (version 2.4.1)[71]. FROGS analysis provides several metrics including the Nearest Sequenced Taxon Index (NSTI) and the percentage of identity and coverage of ASVs on sequences in the PICRUSt2 reference tree. Based on these metrics, we retained only ASVs with NSTI < 0.31, and with an identity and coverage percentage > = 90% (Supplementary Fig. S2). The remaining ASVs were analysed using the complete pipeline with default parameters to obtain information on KEGG function abundance.

Statistical analyses

All statistical analyses were performed using R (version 4.1.3), and the significance threshold was set to 0.05.

Microbiota analyses relied on the R packages Phyloseq[72] (version 1.38), and vegan[73] (version 2.6.2). All analyses except for those of differential abundance were performed on rarefied counts using the smallest sample abundance (i.e., 14,548 sequences). Alpha diversity was estimated with observed richness and the Shannon index, and beta diversity with Bray–Curtis distance. For the beta diversity analysis, because a betadisper test (package vegan) revealed non-homogeneity of dispersion in the line x diet groups, the impacts of line, diet, and their interaction on distance matrices were tested using the anova.cca test (vegan package) on dbRDA results (capscale function from vegan package using Lingoes adjustment). Sample dissimilarities were visualised using a multidimensional scaling (MDS) ordination plot.

For all other univariate analyses (production and efficiency traits, alpha diversities, phylum relative abundance), a linear model with line, diet, and their interaction as main effects was fitted using the lm function, and Wald Chi-square tests for fixed effects were estimated using the Anova function of the package car[74](version

3.1-0). If the interaction was not significant, the model was updated with only the line and the diet as fixed effects; instead, when it was significant, the variance analysis was followed by a post-hoc test (package emmeans[75], version 1.7.5).

For the comparison of SCFA relative concentrations, the same linear model was tested with permutations because of the lack of normality for butyrate, using the lmp function from the lmPerm package[76] (with *perm = "Prob"* and *Ca = 0.01* options) and the anova function, followed by pairwise line x diet group Wilcoxon tests.

To identify ASVs that were differentially abundant between lines in each diet or between diets in each line, we used the DESeq2[77] package (version 1.34) on non-rarefied ASV abundances, with the poscounts estimate size factors method, and Benjamini Hochberg *P*-value adjustment. DESeq2 analyses were also performed on abundance tables of KEGG functions between diets in each line and between lines in each diet, with default parameters but with the more-conservative Benjamini Yekutieli *P*-value adjustment method. Differentially abundant functions were visualised using an Ipath3 metabolic pathways map[78].

Finally, Spearman correlation analysis was performed between the relative concentrations of SCFAs and the relative abundance of bacterial families or genera using the corr.test function from the psych package[79](version 2.2.9) (with *adjust = "BH", use = "complete"* options). Correlation results were visualised using ComplexHeatmap package[80] (version 2.10.0) with default clustering options of rows (families or genera) and column (SCFA), i.e. complete method on Euclidean distance.

Data availability

The sequences generated and analysed during the current study are available in the PRJEB62837 repository, https://www.ebi.ac.uk/ena/browser/view/PRJEB62837.

Received: 20 June 2023; Accepted: 28 March 2024
Published online: 08 April 2024

References

1. OECD-FAO Agricultural Outlook (Edition 2021). OECD https://doi.org/10.1787/4bde2d83-en.
2. Noblet, J., Wu, S.-B. & Choct, M. Methodologies for energy evaluation of pig and poultry feeds: A review. *Anim. Nutr.* **8**, 185–203 (2022).
3. Zerjal, T. *et al.* Assessment of trade-offs between feed efficiency, growth-related traits, and immune activity in experimental lines of layer chickens. *Genet. Sel. Evol.* **53**, 44 (2021).
4. Marchesi, J. A. P. *et al.* Exploring the genetic architecture of feed efficiency traits in chickens. *Sci. Rep.* **11**, 4622 (2021).
5. Gabarrou, J.-F., Geraert, P. A., Williams, J., Ruffier, L. & Rideau, N. Glucose–insulin relationships and thyroid status of cockerels selected for high or low residual food consumption. *Br. J. Nutr.* **83**, 645–651 (2000).
6. de Verdal, H. *et al.* Improving the efficiency of feed utilization in poultry by selection. 1. Genetic parameters of anatomy of the gastro-intestinal tract and digestive efficiency. *BMC Genet.* **12**, 59 (2011).
7. Bindari, Y. R. & Gerber, P. F. Centennial review: Factors affecting the chicken gastrointestinal microbial composition and their association with gut health and productive performance. *Poultry Sci.* **101**, 101612 (2022).
8. Oakley, B. B. *et al.* The chicken gastrointestinal microbiome. *FEMS Microbiol. Lett.* **360**, 100–112 (2014).
9. Kraimi, N. *et al.* Influence of the microbiota-gut-brain axis on behavior and welfare in farm animals: A review. *Physiol. Behav.* **210**, 112658 (2019).
10. Stanley, D. *et al.* Intestinal microbiota associated with differential feed conversion efficiency in chickens. *Appl. Microbiol. Biotechnol.* **96**, 1361–1369 (2012).
11. Yan, W., Sun, C., Yuan, J. & Yang, N. Gut metagenomic analysis reveals prominent roles of Lactobacillus and cecal microbiota in chicken feed efficiency. *Sci. Rep.* **7**, 45308 (2017).
12. Khan, S., Moore, R. J., Stanley, D. & Chousalkar, K. K. The gut microbiota of laying hens and its manipulation with prebiotics and probiotics to enhance gut health and food safety. *Appl. Environ. Microbiol.* https://doi.org/10.1128/AEM.00600-20 (2020).
13. Morrison, D. J. & Preston, T. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut Microbes* **7**, 189–200 (2016).
14. Mahmood, T. & Guo, Y. Dietary fiber and chicken microbiome interaction: Where will it lead to?. *Anim. Nutr.* **6**, 1–8 (2020).
15. Stanley, D. *et al.* Identification of chicken intestinal microbiota correlated with the efficiency of energy extraction from feed. *Vet. Microbiol.* **164**, 85–92 (2013).
16. Stanley, D., Hughes, R. J., Geier, M. S. & Moore, R. J. Bacteria within the gastrointestinal tract microbiota correlated with improved growth and feed conversion: Challenges presented for the identification of performance enhancing probiotic bacteria. *Front. Microbiol.* **7**, 8. https://doi.org/10.3389/fmicb.2016.00187 (2016).
17. Siegerstetter, S.-C. *et al.* Intestinal microbiota profiles associated with low and high residual feed intake in chickens across two geographical locations. *PLOS ONE* **12**, e0187766 (2017).
18. Borey, M. *et al.* Broilers divergently selected for digestibility differ for their digestive microbial ecosystems. *PLOS ONE* **15**, e0232418 (2020).
19. Wen, C. *et al.* Joint contributions of the gut microbiota and host genetics to feed efficiency in chickens. *Microbiome* **9**, 126 (2021).
20. Buzala, M. & Janicki, B. Review: Effects of different growth rates in broiler breeder and layer hens on some productive traits. *Poultry Sci.* **95**, 2151–2159 (2016).
21. Kers, J. G. *et al.* Host and environmental factors affecting the intestinal microbiota in chickens. *Front. Microbiol.* **9**, 322066 (2018).
22. Bordas, A., Tixier-Boichard, M. & Merat, P. Direct and correlated responses to divergent selection for residual food intake in Rhode island red laying hens. *Br. Poult. Sci.* **33**, 741–754 (1992).
23. El-Kazzi, M., Bordas, A., Gandemer, G. & Minvielle, F. Divergent selection for residual food intake in Rhode Island red egg-laying lines: Gross carcase composition, carcase adiposity and lipid contents of tissues. *Br. Poult. Sci.* **36**, 719–728 (1995).
24. Gabarrou, J. F., Géraert, P. A., Picard, M. & Bordas, A. Diet-induced thermogenesis in cockerels is modulated by genetic selection for high or low residual feed intake. *J Nutr* **127**, 2371–2376 (1997).
25. Gabarrou, J. F. *et al.* Energy balance of laying hens selected on residual food consumption. *Br. Poult. Sci.* **39**, 79–89 (1998).
26. Tixier-Boichard, M., Boichard, D., Groeneveld, E. & Bordas, A. Restricted maximum likelihood estimates of genetic parameters of adult male and female Rhode Island red chickens divergently selected for residual feed consumption. *Poult. Sci.* **74**, 1245–1252 (1995).
27. Koh, A., De Vadder, F., Kovatcheva-Datchary, P. & Bäckhed, F. From dietary fiber to host physiology: Short-chain fatty acids as key bacterial metabolites. *Cell* **165**, 1332–1345 (2016).

28. Dong, X. Y., Azzam, M. M. M. & Zou, X. T. Effects of dietary threonine supplementation on intestinal barrier function and gut microbiota of laying hens. *Poult. Sci.* **96**, 3654–3663 (2017).
29. Geng, S. *et al.* Alterations and correlations of the gut microbiome, performance, egg quality, and serum biochemical indexes in laying hens with low-protein amino acid-deficient diets. *ACS Omega* **6**, 13094–13104 (2021).
30. Videnska, P. *et al.* Succession and replacement of bacterial populations in the caecum of egg laying hens over their whole life. *PLOS ONE* **9**, e115142 (2014).
31. Jha, R. & Mishra, P. Dietary fiber in poultry nutrition and their effects on nutrient utilization, performance, gut health, and on the environment: A review. *J. Anim. Sci. Biotechnol.* **12**, 51 (2021).
32. Cantu-Jungles, T. M. & Hamaker, B. R. Tuning expectations to reality: Don't expect increased gut microbiota diversity with dietary fiber. *The Journal of Nutrition* **153**, 3156–3163 (2023).
33. Hamaker, B. R. & Tuncil, Y. E. A perspective on the complexity of dietary fiber structures and their potential effect on the gut microbiota. *Journal of Molecular Biology* **426**, 3838–3850 (2014).
34. Tap, J. *et al.* Gut microbiota richness promotes its stability upon increased dietary fibre intake in healthy adults. *Environ. Microbiol.* **17**, 4954–4964 (2015).
35. Martinez-Guryn, K. *et al.* Small intestine microbiota regulate host digestive and absorptive adaptive responses to dietary lipids. *Cell Host Microbe* **23**, 458-469.e5 (2018).
36. Turnbaugh, P. J. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027–1031 (2006).
37. Velasco-Galilea, M., Piles, M., Ramayo-Caldas, Y. & Sánchez, J. P. The value of gut microbiota to predict feed efficiency and growth of rabbits under different feeding regimes. *Sci. Rep.* **11**, 19495 (2021).
38. Svihus, B. Limitations to wheat starch digestion in growing broiler chickens: A brief review. *Anim. Prod. Sci.* **51**, 583–589 (2011).
39. Tiwari, U. P., Singh, A. K. & Jha, R. Fermentation characteristics of resistant starch, arabinoxylan, and β-glucan and their effects on the gut microbial ecology of pigs: A review. *Anim. Nutr.* **5**, 217–226 (2019).
40. Klostermann, C. E. *et al.* Presence of digestible starch impacts *in vitro* fermentation of resistant starch. *Food Funct.* **15**, 223–235 (2024).
41. Martínez, I., Kim, J., Duffy, P. R., Schlegel, V. L. & Walter, J. Resistant starches types 2 and 4 have differential effects on the composition of the fecal microbiota in human subjects. *PLoS ONE* **5**, e15046 (2010).
42. Regmi, P. R., Metzler-Zebeli, B. U., Gänzle, M. G., van Kempen, T. A. T. G. & Zijlstra, R. T. Starch with high amylose content and low in vitro digestibility increases intestinal nutrient flow and microbial fermentation and selectively promotes bifidobacteria in pigs. *J. Nutr.* **141**, 1273–1280 (2011).
43. Ryan, S. M., Fitzgerald, G. F. & van Sinderen, D. Screening for and identification of starch-, amylopectin-, and pullulan-degrading activities in bifidobacterial strains. *Appl. Environ. Microbiol.* **72**, 5289–5296 (2006).
44. Rivière, A., Selak, M., Lantin, D., Leroy, F. & De Vuyst, L. Bifidobacteria and butyrate-producing colon bacteria: Importance and strategies for their stimulation in the human gut. *Front. Microbiol.* https://doi.org/10.3389/fmicb.2016.00979 (2016).
45. Takada, T., Kurakawa, T., Tsuji, H. & Nomoto, K. Fusicatenibacter saccharivorans gen. nov., sp. Nov., isolated from human faeces. *Int. J. Syst. Evol. Microbiol.* **63**, 3691–3696 (2013).
46. Wongkuna, S. *et al.* Taxono-genomics description of Olsenella lakotia SW165T sp. nov., a new anaerobic bacterium isolated from cecum of feral chicken. *F1000Res* **9**, 1103 (2020).
47. Lundberg, R., Scharch, C. & Sandvang, D. The link between broiler flock heterogeneity and cecal microbiome composition. *Anim. Microbiome* **3**, 54 (2021).
48. Zhang, Y. *et al.* Dietary resistant starch modifies the composition and function of caecal microbiota of broilers. *J. Sci. Food Agric.* **100**, 1274–1284 (2020).
49. Holmstrøm, K., Collins, M. D., Møller, T., Falsen, E. & Lawson, P. A. Subdoligranulum variabile gen. nov., sp. nov. from human feces. *Anaerobe* **10**, 197–203 (2004).
50. Khan, M. T. *et al.* The gut anaerobe *Faecalibacterium prausnitzii* uses an extracellular electron shuttle to grow at oxic–anoxic interphases. *ISME J.* **6**, 1578–1585 (2012).
51. Moens, F., Weckx, S. & De Vuyst, L. Bifidobacterial inulin-type fructan degradation capacity determines cross-feeding interactions between bifidobacteria and Faecalibacterium prausnitzii. *Int. J. Food Microbiol.* **231**, 76–85 (2016).
52. Ziemer, C. J. Newly cultured bacteria with broad diversity isolated from eight-week continuous culture enrichments of cow feces on complex polysaccharides. *Appl. Environ. Microbiol.* **80**, 574–585 (2014).
53. Zhou, Q. *et al.* Genetic and microbiome analysis of feed efficiency in laying hens. *Poult. Sci.* **102**, 102393. https://doi.org/10.1016/j.psj.2022.102393 (2022).
54. Zhang, Y. K. *et al.* Characterization of the rumen microbiota and its relationship with residual feed intake in sheep. *Animal* **15**, 100161 (2021).
55. Louis, P. & Flint, H. J. Formation of propionate and butyrate by the human colonic microbiota. *Environ. Microbiol.* **19**, 29–41 (2017).
56. Torok, V. A. *et al.* Identification and characterization of potential performance-related gut microbiotas in broiler chickens across various feeding trials. *Appl. Environ. Microbiol.* **77**, 5868–5878 (2011).
57. Singh, K. M. *et al.* High through put 16S rRNA gene-based pyrosequencing analysis of the fecal microbiota of high FCR and low FCR broiler growers. *Mol Biol Rep* **39**, 10595–10602 (2012).
58. Gardiner, G. E., Metzler-Zebeli, B. U. & Lawlor, P. G. Impact of intestinal microbiota on growth and feed efficiency in pigs: a review. *Microorganisms* **8**, 1886 (2020).
59. De Maesschalck, C. *et al.* Amorphous cellulose feed supplement alters the broiler caecal microbiome. *Poult. Sci.* **98**, 3811–3817 (2019).
60. Polansky, O. *et al.* Important metabolic pathways and biological processes expressed by chicken cecal microbiota. *Appl. Environ. Microbiol.* **82**, 1569–1576 (2016).
61. den Besten, G. *et al.* The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. *J. Lip. Res.* **54**, 2325–2340 (2013).
62. Van Soest, P. J., Robertson, J. B. & Lewis, B. A. Methods for dietary fiber, neutral detergent fiber, and nonstarch polysaccharides in relation to animal nutrition. *J. Dairy Sci.* **74**, 3583–3597 (1991).
63. Bedu-Ferrari, C. *et al.* In-depth characterization of a selection of gut commensal bacteria reveals their functional capacities to metabolize dietary carbohydrates with prebiotic potential. *Systems* https://doi.org/10.1128/msystems.01401-23 (2024).
64. Godon, J. J., Zumstein, E., Dabert, P., Habouzit, F. & Moletta, R. Molecular microbial diversity of an anaerobic digestor as determined by small-subunit rDNA sequence analysis. *Appl. Environ. Microbiol.* **63**, 2802–2813 (1997).
65. Lluch, J. *et al.* The characterization of novel tissue microbiota using an optimized 16S metagenomic sequencing pipeline. *PLoS ONE* **10**, e0142334 (2015).
66. Nadkarni, M. A., Martin, F. E., Jacques, N. A. & Hunter, N. Determination of bacterial load by real-time PCR using a broad-range (universal) probe and primers set. *Microbiology* **148**, 257–266 (2002).
67. Escudié, F. *et al.* FROGS: find, rapidly, OTUs with galaxy solution. *Bioinformatics* **34**, 1287–1294 (2018).
68. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: A fast and accurate illumina paired-end reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).

69. Bokulich, N. A. *et al.* Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* **10**, 57–59 (2013).
70. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2012).
71. Douglas, G. M. *et al.* PICRUSt2: An improved and extensible approach for metagenome inference. *bioRxiv* https://doi.org/10.1101/672295 (2019).
72. McMurdie, P. J. & Holmes, S. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **8**, e61217 (2013).
73. Oksanen, J. *et al.* Vegan: Community ecology package (2022).
74. Fox, J. *et al.* Car: Companion to Applied Regression (2022).
75. Lenth, R. V. *et al.* Emmeans: Estimated marginal means, aka least-squares means (2022).
76. Wheeler, B. & Torchiano, M. lmPerm: permutation tests for linear models (2016).
77. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
78. Darzi, Y., Letunic, I., Bork, P. & Yamada, T. iPath30: Interactive pathways explorer v3. *Nucleic Acids Res.* **46**, 510–513 (2018).
79. Revelle, W. psych: Procedures for psychological, psychometric, and personality research (2023).
80. Gu, Z. Complex heatmap visualization. *iMeta* **1**, e43 (2022).

Acknowledgements

The data were collected in the framework of projects that received financial support from the French National Research Agency (ChickStress project, ANR-13-ADAP) and from the European Union's H2020 programme under Grant Agreement No. 633531 (Feed-a-Gene project). The authors thank the staff of the INRAE experimental poultry unit (UE1295 PEAT, Nouzilly, France) for producing and rearing the animals. This sequencing experiment was performed in collaboration with the GeT core facility, Toulouse, France (GeT, https://doi.org/https://doi.org/10.15454/1.5572370921303193E12), and was supported by the national infrastructure programme France Génomique, funded as part of the "Investissement d'avenir" programme managed by the Agence Nationale pour la Recherche (contract ANR-10-INBS-09). We are also grateful to the GenoToul bioinformatics platform Toulouse Occitanie for providing computing and storage resources (Bioinfo Genotoul, https://doi.org/https://doi.org/10.15454/1.5572369328961167E12). Finally, we thank Andrea Rau, Denis Laloë, and Mahendra Mariadassou for providing valuable suggestions on statistical analyses.

Author contributions

T.Z. and S.L. conceived the idea, obtained funding, and participated in sample collection. D.J. carried out DNA extractions, N.B. participated in the 16S amplification, and J.-L.C. contributed to SCFA quantification. M.B. carried out all analyses and A.L. participated in the bioinformatics analysis. M.B. and T.Z. wrote the manuscript. F.C. and G.P. helped with interpretation of the data and, together with A.M., A.L. and S.L., contributed useful comments and suggestions on the manuscript draft. All authors read and approved the final version.

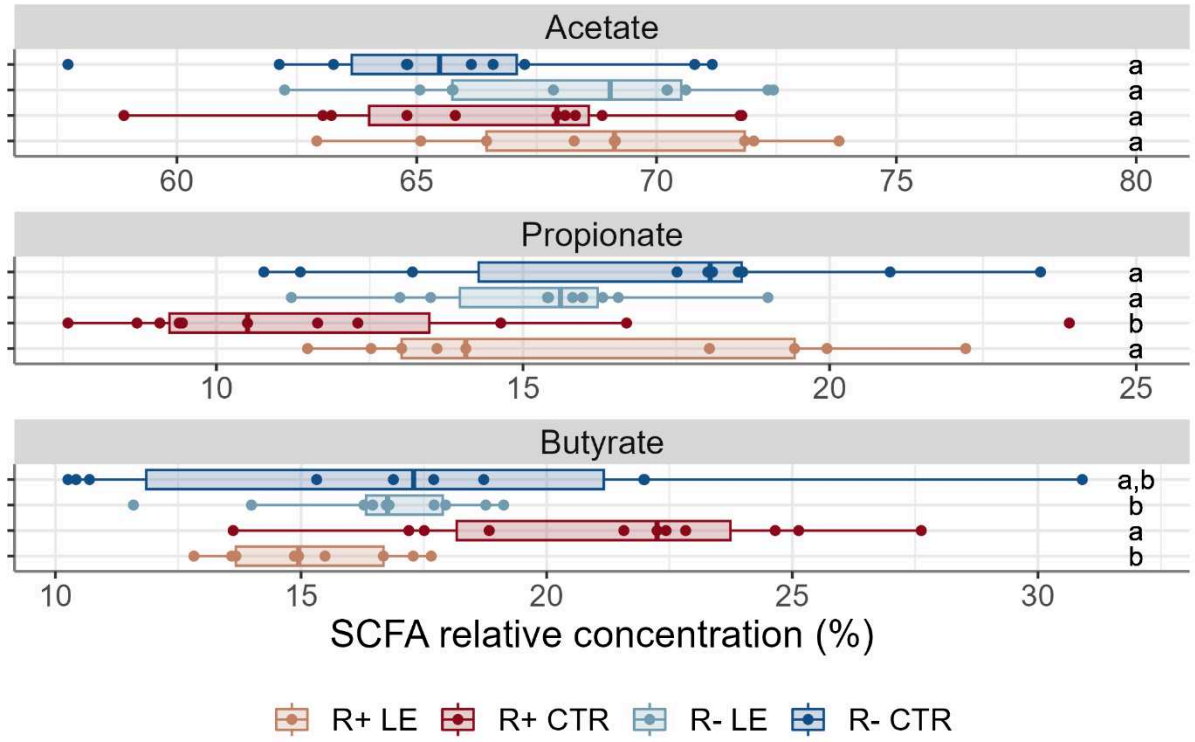
Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-58374-3.

Correspondence and requests for materials should be addressed to M.B. or T.Z.



Supplementary Figure S1 : Distribution of relative SCFA concentration. Distribution of the relative SCFA concentration (%) within each line x diet group: red= R+ line fed the CTR diet; blue= R- line fed the CTR diet; orange= R+ line fed the LE diet; light blue= R- line fed the LE diet. Different letters at the top indicate significant differences from Wilcoxon pairwise tests;

E. Partie 2 : Constitution d'un catalogue enrichi de taxonomies et de fonctions

E.1. Contexte et objectifs

L'analyse de données de séquences de métabarcoding (chapitre précédent **D**) nous a permis de poser des hypothèses biologiques quant au rôle du microbiote cæcal sur l'efficience alimentaire des poules pondeuses, et l'influence du régime sur la composition microbienne et fonctionnelle de ce microbiote. En revanche, en se basant sur le séquençage uniquement de la région V3-V4 du gène de l'ARNr 16S, la résolution taxonomique obtenue sur cette première analyse n'est que de 20% à l'espèce et 60% au genre. Comme indiqué en introduction (paragraphe **A.5.1.d**), cette faible résolution taxonomique peut être due à une capacité incomplète de la région ciblée à discriminer l'ensemble des organismes présents et/ou au fait que la référence taxonomique utilisée (la base SILVA 138.1) ne soit pas complète ou parfaitement annotée taxonomiquement. Par ailleurs, l'incomplétude des bases de référence a également été observée au niveau de l'analyse par inférence des fonctions. En effet, celle-ci a été menée seulement sur une partie des ASV (bien que majoritaire) et sans assurance que les génomes de référence les plus proches identifiés et utilisés pour inférer les fonctions soient effectivement ceux dont sont issus ces ASV. Ainsi, ces premiers résultats d'analyse de données de métabarcoding ont des lacunes que l'analyse de données de séquences métagénomiques peut contribuer à réduire.

Lors du séquençage complet de l'ADN, tous les gènes de tous les organismes sont séquencés. Par analyse bioinformatique, cet ensemble de lectures permettra de reconstruire des gènes que l'on pourra annoter fonctionnellement, ainsi que des génomes des différents organismes ou plus exactement des espèces métagénomiques (MGS), que l'on pourra affilier taxonomiquement. Comme indiqué dans le paragraphe **C.3.2.a**, la production de ce séquençage complet de l'ADN fait partie intégrante du projet MetaChick. Ce projet a notamment produit en amont de cette thèse, un catalogue de référence de gènes et de MGS à partir d'échantillons de contenu cæcal d'une diversité importante de poulets de chair et de poules pondeuses français. Ce catalogue inclut, en particulier, 20 échantillons de notre condition CTR (10 par lignée divergente pour l'efficience alimentaire, R+ et R-). Bien que l'analyse qualitative préliminaire de ce séquençage (paragraphe **C.3.4.b**) ait montré un très bon taux d'alignement des lectures sur ce catalogue MetaChick pour l'ensemble de nos échantillons, nous nous sommes demandé ce que pourraient apporter au catalogue MetaChick les échantillons LE provenant de poules nourries avec un régime non conventionnel, l'alimentation étant l'un des facteurs les plus importants de structuration du microbiote.

Pour l'analyse *de novo* des séquences, il est en général recommandé d'appliquer la chaîne de traitement – assemblage des lectures en contigs, annotation structurale des gènes, « *binning* » des contigs en MAG* –, échantillon par échantillon (**Figure E-1**).

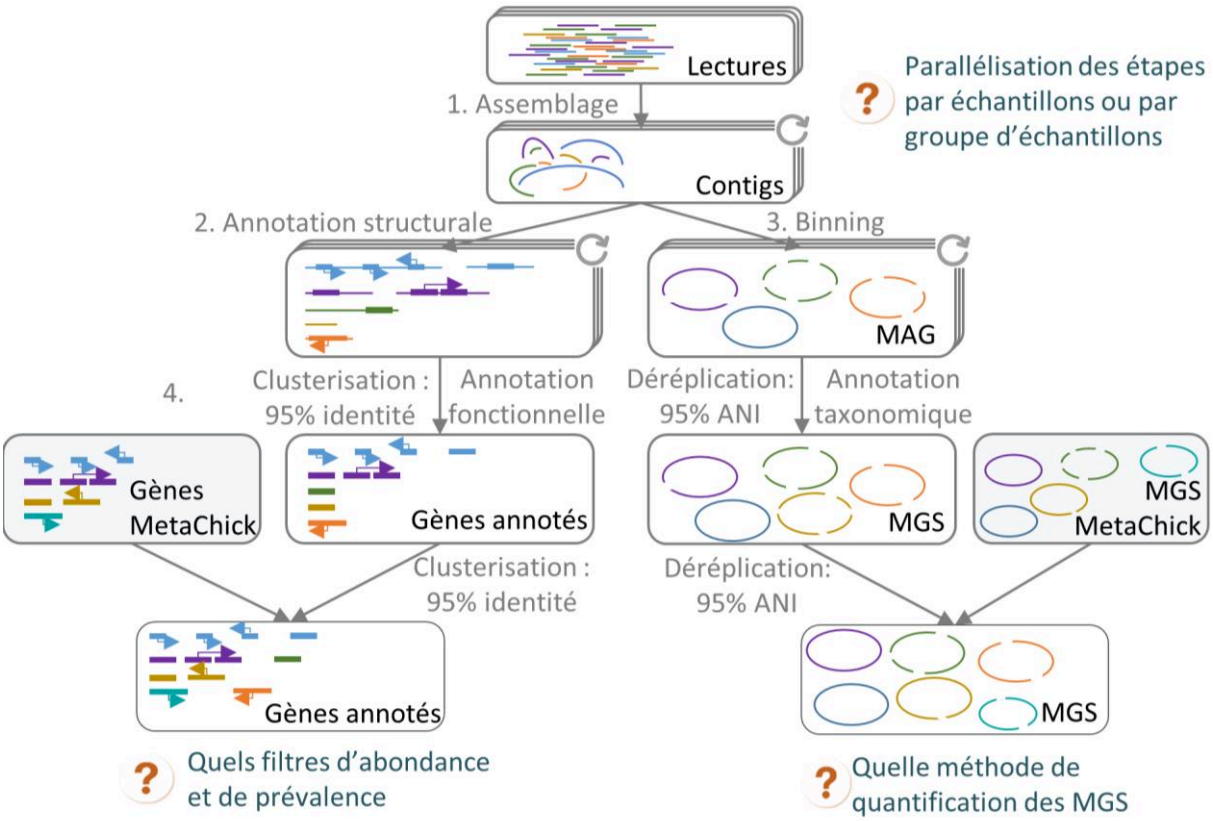


Figure E-1: Schéma du processus de traitement bioinformatique *de novo* des séquences issues de séquençage métagénomique avec intégration d'une référence externe.
Les flèches circulaires indiquent les étapes que l'on recommande de faire par échantillon. Les points d'interrogation reflètent les réflexions sur la stratégie d'application de ce processus et des méthodes de quantification.

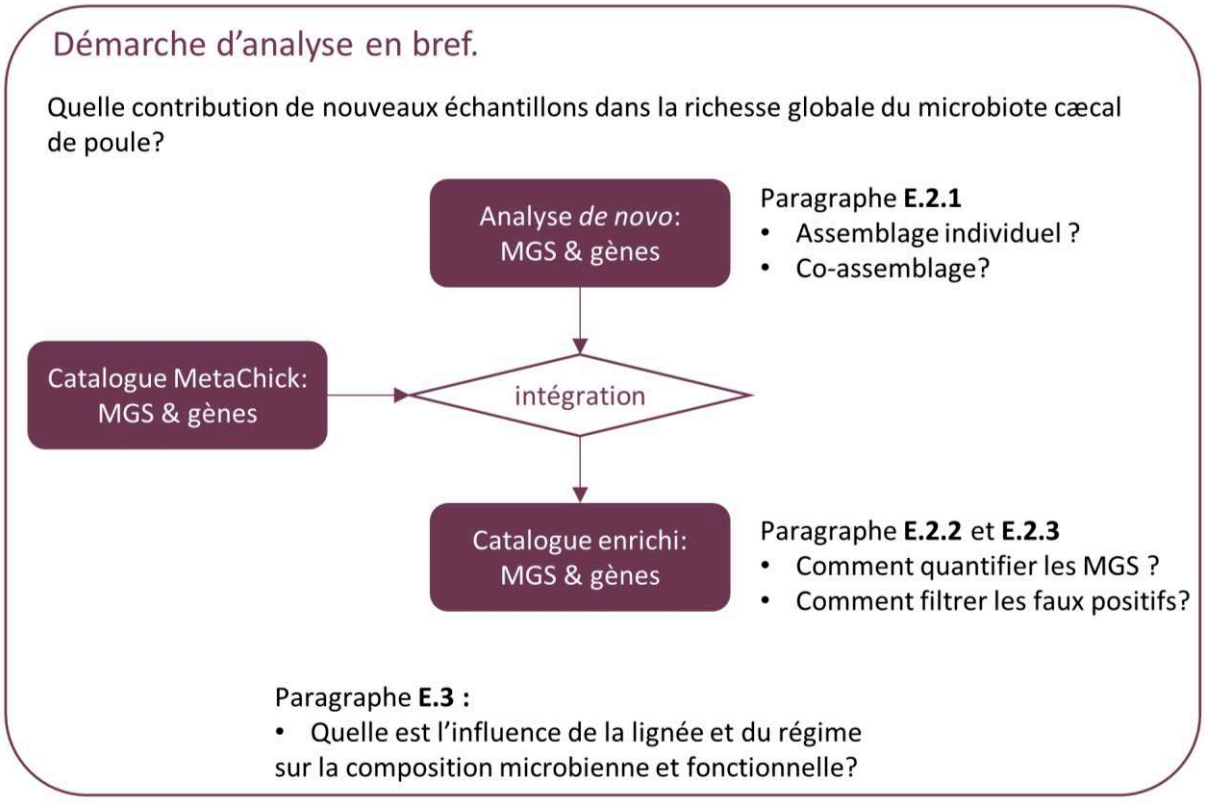
En effet, les écosystèmes microbiens, en particulier intestinaux, sont souvent composés de micro-organismes avec de forts déséquilibres d'abondances, ce qui complexifie l'assemblage des lectures. Mélanger des échantillons (dans une stratégie de co-assemblage) permet d'augmenter le nombre de lectures provenant notamment d'espèces rares, mais si la proportion des micro-organismes dans les différents échantillons mélangés est variable alors la complexité d'assemblage est encore augmentée. L'inconvénient de la stratégie par échantillon est qu'elle nécessite une profondeur plus importante de séquençage pour chaque échantillon dans le but d'assurer la représentation de l'ensemble des micro-

* MAG : « *Metagenome-Assembled Genome* », assemblage d'un métagénome représentant théoriquement une entité taxonomique. Leur déréplication à 95% d'ANI produit des MGS (« *Metagenomic Species* »), représentant théoriquement les espèces.

organismes même en faible abondance. Or, dans la dernière phase de séquençage du projet MetaChick, les échantillons (dont nos échantillons de la condition LE), ont été séquencés deux fois moins profondément que les échantillons de phase 1 du projet, échantillons qui ont servi à l'établissement du catalogue (dont nos échantillons CTR) (paragraphe **C.3.4.b** et **Figure C-9**). Avant d'évaluer la contribution que pourraient apporter nos échantillons LE au catalogue MetaChick, j'ai donc procédé à une comparaison des stratégie d'analyse *de novo*, soit en l'appliquant à l'échelle de chaque échantillon (assemblage individuel) ou bien à l'échelle d'un groupe d'échantillons provenant de la combinaison d'une lignée et d'un régime (co-assemblage par « lignée x régime ») en faisant l'hypothèse que les abondances des micro-organismes étaient relativement stables au sein de chaque groupe. Les résultats de cette première analyse sont décrits dans la section suivante **E.2.1**.

Après l'étape d'assemblage des lectures, d'annotation structurale et de « *binning* », la redondance des gènes et des MGS entre échantillons ou groupes d'échantillons (« lignée x régime ») dans le cas des co-assemblages, doit être réduite (étape 4 de la **Figure E-1**). Par ailleurs, pour assurer une complétude la plus importante possible de la diversité des gènes et MGS présents dans nos échantillons, nous avons enrichi (par déréplication des MGS et clusterisation des gènes) les résultats obtenus *de novo* avec le catalogue MetaChick. Nous aurions pu choisir un catalogue comme le GG-IGC (Feng et al. 2021) qui intègre des échantillons de contenu cæcal et d'autres compartiments intestinaux, majoritairement de Chine mais également d'Europe. La diversité importante de ces échantillons permet théoriquement de couvrir une large diversité microbienne du microbiote intestinal. Cependant, une partie de cette diversité n'est vraisemblablement pas représentative de nos échantillons cæcaux de poule pondeuse élevées en France. Par ailleurs, les analyses comparatives menées sur le catalogue MetaChick ont montré qu'une part importante des MGS étaient spécifiques à l'un ou l'autre de ces catalogues (résultats non publiés). Nous avons donc fait le choix de prendre en compte le catalogue MetaChick qui restreint la diversité des échantillons à des échantillons exclusivement de contenu cæcal, d'animaux élevés dans des environnements proches de nos échantillons. Cela nous permet d'enrichir notre analyse *de novo* tout en limitant le bruit qui augmenterait la détection de faux positifs. La fusion de deux jeux de données nous a amenés à réfléchir à deux autres questions méthodologiques : i) quelle méthode de quantification des MGS doit être utilisée: une quantification basée sur l'abondance moyenne des gènes marqueurs ou bien une quantification basée sur la profondeur de réalignement des lectures ? et ii) quel seuil d'abondance et de prévalence doit être utilisé pour filtrer les gènes et les MGS ? Le catalogue MetaChick fournit également pour chaque MGS, un pan-génome d'espèce métagénomique (MSP), autrement dit la composition en gènes et notamment en gènes marqueurs de chaque génome représentant les différentes espèces. Dans notre cas, après intégration des deux analyses (MetaChick et *de novo*), les gènes marqueurs doivent être identifiés sur les potentielles

nouvelles MGS détectées *de novo*. Or, nous travaillons ici sur un petit nombre d'échantillons avec des profondeurs de séquençage variable (Plaza Oñate et al. 2019). Par ailleurs, bien que l'intégration du catalogue MetaChick permette de compléter les résultats obtenus *de novo*, elle ajoute nécessairement un certain nombre de gènes et de MGS qui ne sont pas présents dans nos échantillons. Identifier ces gènes et MGS en fonction de leur profil d'abondance et de prévalence permettra de limiter le nombre de faux positifs dans la suite de l'analyse. Les sections **E.2.2** et **E.2.3** présente les résultats de l'intégration des deux catalogues, la comparaison entre les méthodes de quantification des MGS, ainsi que la calibration des filtres d'abondance et de prévalence pour limiter les faux positifs. Enfin, la section **E.3** présente le catalogue enrichi obtenu en termes de composition taxonomique, de fonctions et de sa représentation dans nos deux lignées divergentes pour l'efficience alimentaire nourries avec deux régimes alimentaires.



E.2. Mise au point méthodologique d'analyse de novo et d'intégration du catalogue de gènes et de MGS MetaChick

E.2.1. Comparaison de stratégies d'analyse de novo par assemblage individuel ou co-assemblage

Que ce soit pour la stratégie d'assemblage individuel ou pour la stratégie de co-assemblage, l'analyse s'est appuyée sur la chaîne de traitement metagWGS (l'enchaînement des outils et le paramétrage du pipeline sont développés dans le chapitre précédent, paragraphe **C.3.2.b** et **Figure C-6**). Pour comparer les deux stratégies d'analyses (assemblage individuel ou co-assemblage), nous nous sommes appuyés sur des mesures quantitatives et qualitatives à l'échelle des contigs, des gènes et des MGS.

Comparaison des stratégies d'assemblage à l'échelle des contigs

La qualité d'un assemblage se mesure selon le nombre de contigs générés et leur longueur. Plus un assemblage contient un faible nombre de contigs avec une majorité de grande taille, meilleure sera sa qualité (**Tableau E-1**). Une des mesures de longueur que l'on souhaite augmenter est le N50. Il correspond à la longueur minimum des plus longs contigs incluant 50% de la taille cumulée de l'assemblage. A taille cumulée constante, plus le N50 est grand et moins l'assemblage est considéré comme fragmenté.

Tableau E-1 : Mesures quantitatives et qualitatives des contigs.

		Assemblage individuel	**Co-assemblage**
Tous les contigs	Nombre	1 152 567,82	2 595 125
	Longueur N50 (pb)	1 167,51	2 935,7
	Longueur cumulée (Mb)	716,4	2 097,7
Contigs > 1 kb	Nombre et pourcentage	103 429,8 (9.0%)	332 979,2 (12,83%)
	Longueur N50 (pb)	5 635,51	8 172,7
	Longueur cumulée (Mb) et pourcentage	351,4 (49%)	1 381,4 (65,8%)
	Taux de lectures proprement ré-alignées	77,10%	80,10%

Les assemblages étant générés par individus (39 assemblages individuels) ou par groupe d'échantillons « lignée x régime » (4 co-assemblages), les mesures reportées sont des moyennes.

A partir de ces premières mesures, les co-assemblages présentent de meilleurs résultats. Bien que générant plus de 2 fois plus de contigs pour une longueur cumulée presque 3 fois plus grande que les assemblages individuels, ils présentent également une valeur de N50 2,5 fois plus importante. Sans être moins fragmentés, on peut supposer que les co-assemblages sont plus complets. En cohérence avec ces premières observations, le filtre sur une taille minimale de 1kb est moins impactant sur les contigs obtenus par co-assemblage que sur ceux obtenus par assemblage individuel et permet de conserver en moyenne 65,8% des assemblages de chaque groupe d'échantillons, et d'améliorer le taux moyen de réalignement des lectures de 3%.

Ces mesures sont toutefois variables selon les groupes d'échantillons « lignée x régime » (**Tableau E-2**). Pour rappel, les échantillons des groupes CTR ont été séquencés à une profondeur deux fois plus importante que les échantillons des groupes LE. Nous nous attendions à ce que cette différence de profondeur impacte les qualités d'assemblage et en effet, on observe que les tailles de N50 sont près de deux fois plus petites pour les assemblages des échantillons de la condition LE que pour ceux de la condition CTR. On observe aussi une plus forte diminution de la taille cumulée des contigs de tailles supérieures à 1kb pour les échantillons LE que pour les échantillons CTR. Cette diminution est d'autant plus vraie pour les assemblages individuels que pour les co-assemblages.

Tableau E-2 : Mesures de longueurs d'assemblage par groupe lignée x régime.

		R- CTR	R+ CTR	R- LE	R+ LE
Assemblage individuel	**Tous les contigs**				
	Longueur cumulée (Mb)	767,8	682,1	682,8	734,7
	N50	1 387,6	1 927,7	649,3	654,1
	Contigs > 1kb				
	Longueur cumulée (Mb et %)	431,8 / 56%	410,3 / 60%	268,9 / 39%	288,3 / 39%
	N50	6 255,6	8 726,2	3 802,7	3 548,8
Co-assemblage	**Tous les contigs**				
	Longueur cumulée (Mb)	2 087,3	1 935,9	2 187,9	2 179,6
	N50	4 053	4 077	1 786	1 827
	Contigs > 1kb				
	Longueur cumulée (Mb et %)	1 482,2 / 71%	1 366,4 / 70%	1 338,5 / 61%	1 338,2 / 61%
	N50	10 136	10 682	5 929	5 944

Pour les assemblages individuels, les valeurs correspondent aux moyennes sur ~10 échantillons d'un groupe donné et pour les co-assemblages les valeurs sont les mesures exactes pour chaque groupe.

Comparaison des stratégies d'assemblage à l'échelle de l'annotation structurale et fonctionnelle

A l'image de la comparaison des contigs, l'annotation structurale[*] est plus efficace lorsqu'elle est basée sur les contigs issus du co-assemblage (**Tableau E-3**). En effet, il a été détecté davantage d'éléments génomiques (avec une quasi-dominance (> 99%) de gènes potentiellement codant des protéines *i.e.* des CDS), et de tailles globalement plus grandes. Au niveau de l'annotation fonctionnelle[**], 6 435 fonctions KEGG ont été détectées dans les deux analyses, et 106 et 219 sont respectivement spécifiques de l'analyse par assemblage individuel et de celle par co-assemblage. A l'échelle des voies métaboliques, 2 voies sont détectées uniquement avec l'analyse par assemblage individuel, 26 uniquement avec l'analyse par co-assemblage, et 389 sont détectées en commun.

Tableau E-3 : Résumé quantitatif de l'annotation structurale et fonctionnelle.

		Assemblage individuel	**Co-assemblage**
Nombre / pourcentage	Eléments	2 102 272	2 255 430
	CDS	2 092 440 (99,5%)	2 244 999 (99,5%)
Longueur	Cumulée (Mb)	1 685,6	1 887,5
	Minimum	52	53
	Maximum	34 467	42 309
	Moyenne	801,8	836,9
	N50	1 029	1 074
Nombre et pourcentage en fonction de la longueur	< 700 pb	1 109 971 (52,8%)	1 121 407 (49,7%)
	>= 700 pb & <= 2 kb	906 843 (43,1%)	1 031 442 (45,7%)
	> 2kb	85 458 (4%)	102 581 (4,5%)
Annotation KEGG	Eléments annotés	853 913 (40,6%)	984 934 (43,7%)
	Annotations	6 541 fonctions, 365 voies métaboliques	6 654 fonctions, 389 voies métaboliques
	Taux de lectures proprement ré-alignées	55,04%	58,01%

Comparaison des stratégies d'assemblage à l'échelle des MGS

Chaque MGS obtenue après déréplication à 95% d'ANI est un ensemble non ordonné de contigs représentant théoriquement une espèce. Pour évaluer la qualité de chaque MGS, l'outil CheckM2 (Chklovski et al. 2022) utilise un jeu de séquences de gènes universels attendus en simple copie dans

[*] l'annotation structurale correspond à l'identification des structures génomiques (CDS, ARNr, ARNt,…)
[**] l'annotation fonctionnelle correspond à l'attribution de fonction aux gènes

les génomes, et calcule la complétude* et la contamination**. En fonction de ces deux mesures, il est possible de classer les MGS en deux catégories : (i) les MGS de haute qualité (HQ) sont complètes à plus de 90% et contiennent moins de 5% de contamination et (ii) les MGS de qualité moyenne (MQ), sont complètes à plus de 50% et contiennent moins de 10% de contamination. Dans notre comparaison, les co-assemblages permettent de reconstruire quelques MGS en plus dont 6 de haute qualité et 30 de moyenne qualité (**Tableau E-4**). Ils intègrent plus de contigs dont le N50 est significativement plus grand (*P*-value < 0.001). De plus, bien que la majorité des MGS supplémentaires soient de qualité moyenne, la complétude est similaire (en moyenne à 83%) et la contamination est significativement plus faible que pour les MGS obtenues sur les assemblages individuels (en moyenne à 0.91% contre 1,25%, *P*-value < 0.001).

Du point de vue des taxonomies associées à ces MGS, l'analyse faite par co-assemblage est moins précise avec notamment plus de MGS affiliées jusqu'au genre et moins jusqu'à l'espèce (282 contre 288 pour les MGS issues des assemblages individuels). Pour autant, les co-assemblages permettent d'identifier plus de taxonomies différentes et ce à chaque rang taxonomique de la classe à l'espèce. Ceci s'explique par la construction de plusieurs MGS pour une même espèce, représentant potentiellement des souches différentes. Cette redondance des taxonomies est moins importante parmi les MGS obtenues par co-assemblage (281 espèces sur 282 MGS affiliées à l'espèce) que parmi les MGS obtenues par assemblage individuel (242 espèces sur 288 MGS affiliées à l'espèce). Par ailleurs, l'ensemble des taxonomies jusqu'au genre détectées *via* les assemblages individuels sont bien retrouvées grâce aux co-assemblages. A l'échelle des espèces, 225 espèces sont détectées dans les deux analyses, 17 ne sont détectées que *via* les assemblages individuels et 56 sont détectées uniquement *via* les co-assemblages. Concernant les 17 espèces non représentées dans les co-assemblages, 16 le sont parmi les MGS du catalogue MetaChick. La dernière espèce *Onthousia excrementipullorum* est représentée par une MGS assemblée à partir de séquences provenant de deux échantillons CTR. *O. excrementipullorum* est représentée dans la base GTDB par des métagénomes provenant également d'échantillons de contenu cæcal de poule (Glendinning et al. 2020; Gilroy et al. 2021). Il est donc probable que l'analyse par co-assemblage ainsi que le catalogue MetaChick ne soient pas exhaustifs du point de vue de la diversité du microbiote cæcal de nos échantillons.

* Complétude : pourcentage de gènes universels retrouvés d'une MGS
** Contamination : pourcentage de gènes attendus en simple copie et présents en plusieurs copies.

Tableau E-4 : Mesure quantitative et qualitative des MGS.

		Assemblage individuel	**Co-assemblage**
Nombre et qualité		566 dont 261 HQ, 305 MQ	603 dont 267 HQ, 335 MQ
Proportion de contigs > 1kb inclus		27%	20,3%
Longueur	Minimum (kb)	453,8	508,1
	Moyenne (kb)	1 900	1 931,5
	Maximum (kb)	5 676,4	5 960,3
	Cumulée (Mb)	1 075,4	1 164,7
N50 des contigs par MGS* **	Minimum	1 300	1 681
	Moyenne	38 799,51	40 350,28
	Maximum	1 107 164	1 111 641
Complétude	Minimum	50,06	50,01
	Moyenne	82,37	83,43
	Médiane	88,11	87,69
	Maximum	100	100
Contamination* **	Minimum	0	0
	Moyenne	1,25	0,91
	Médiane	0,73	0,56
	Maximum	8,67	11,87
Taxonomies	**MGS annotées jusqu'à**		
	Domaine	1 (0,2%)	2 (0,33%)
	Famille	2 (0,3%)	1 (0,17%)
	Genre	275 (48,6%)	317 (52,57%)
	Espèce	288 (50,9%)	282 (46,77%)

	individuel spécifique	**intersection**	**co-assemblage spécifique**
Nombre de taxons			
Domaine	0	2	0
Phylum	0	13	0
Classe	0	15	1
Ordre	0	27	6
Famille	0	61	8
Genre	0	211	67
Espèce	17	225	56

*** indique un test de Wilcoxon significatif avec une *P*-value < 0.001.

Au vu de l'ensemble de ces comparaisons, la stratégie de co-assemblage semble répondre à la problématique de la plus faible profondeur de séquençage des échantillons de la condition de régime alimentaire LE. Cette stratégie permet une plus grande détection de gènes et de fonctions, une

reconstruction de plus nombreuses MGS et l'identification d'un plus grand nombre d'espèces. Pour pallier son défaut de pouvoir discriminant qui lui fait perdre certains gènes et certaines espèces (ou potentiellement certaines souches), deux stratégies peuvent être mise en place : une analyse hybride combinant les résultats d'analyse individuelle et co-assemblage, ou l'utilisation d'un catalogue de référence. Pour notre étude, nous avons choisi de poursuivre l'analyse du microbiote cæcal en utilisant les résultats des co-assemblages enrichis par ceux du catalogue MetaChick.

E.2.2. Intégration du catalogue MetaChick et des analyses de novo de co-assemblage

Bénéficier du catalogue MetaChick obtenu sur des écosystèmes cæcaux proches de nos échantillons est un avantage pour couvrir au mieux la diversité fonctionnelle et taxonomique. Après l'annotation fonctionnelle sur la base de données eggNOG 5.0 et la réannotation taxonomique sur la base GTDB214, le catalogue MetaChick et les résultats obtenus sur les co-assemblages ont été intégrés en répétant les étapes de réduction de la redondance, *i.e* clusterisation des gènes à un seuil d'identité de 95% et déréplication des MGS à un seuil ANI de 95% (**Figure E-1**).

Au niveau des gènes, l'intégration des deux catalogues a un effet modéré par rapport au contenu fonctionnel du catalogue MetaChick (**Tableau E-5**). Le catalogue enrichi contient 13,76 millions de gènes dont 1,4% de nouveaux gènes issus des co-assemblages, mais le nombre de gènes annotés est moins important que dans le catalogue MetaChick. Il contient 13 fonctions KEGG de plus mais en perd 3, et ces nouvelles annotations ne permettent pas de détecter de nouvelles voies métaboliques.

Au niveau des MGS, l'impact de l'intégration est légèrement plus significatif. Les co-assemblages permettent d'ajouter 41 nouvelles MGS dont 6 de haute qualité, et 144 autres MGS provenant des co-assemblages sont retenues comme metagénomes représentatifs, *i.e.* considérés de meilleure qualité du point de vue notamment de la complétude et de la contamination. Enfin, davantage de MGS sont assignées jusqu'à l'espèce et toutes les taxonomies sauf une espèce (*Phocaeicola dorei*) du catalogue MetaChick sont bien représentées dans le catalogue enrichi. Il contient notamment 8 espèces et 7 genres non détectés initialement dans le catalogue MetaChick. En ce qui concerne *P. dorei*, la MGS du catalogue MetaChick la représentant se retrouve agrégée, dans le catalogue enrichi, avec une autre MGS du catalogue MetaChick et une MGS issue des co-assemblages *de novo*. Ces deux MGS sont affiliées à une seconde espèce du même genre, *Phocaeicola vulgatus*, partageant plus de 95,2% d'ANI avec *P. dorei* selon la base GTDB. Tout comme initialement en analyse métabarcoding du gène de l'ARNr 16S, l'application d'un seuil fixe pour différencier les espèces ne semble pas systématiquement pertinent. Des traitements *a posteriori* de la déréplication des MGS (dans le cas de l'intégration d'un catalogue extérieur) ou des MAG (dans le cas d'une analyse *de novo*) pourraient être appliqués pour

tenter de mieux discriminer ces espèces proches. Par exemple, lors de la création du catalogue MetaChick, l'identification et l'analyse des profils d'abondances des gènes marqueurs et accessoires (utilisés pour discriminer des souches) a initialement permis de séparer ces deux espèces de *Phocaeicola*. Il pourrait également s'agir d'un contrôle des taxonomies de chaque MAG composant la MGS finale, combiné à des déréplications à des seuils ANI plus haut (le seuil de 99% d'ANI est communément utilisé pour différencier les souches).

Tableau E-5 : Comparaison du catalogue MetaChick et du catalogue *de novo* enrichi.

		Catalogue MetaChick	**Catalogue enrichi**
Gènes	**Nombre de gènes**	13 627 403	13 765 236
	Nouveaux gènes		201 307
	Gènes annotés (KEGG)	5 008 124 (36,8%)	4 298 329 (31,2%)
	Nouveaux gènes annotés		51 029
	Annotations KEGG	10 907 KO, 425 pathways	10 917 KO, 425 pathways
	Annotations KEGG supplémentaires / perdues		+13 KO / -3 KO
MGS	**Nombre et quantité**	2 629 dont 1 649 HQ, 976 MQ	2 668 dont 1 676 HQ, 988 MQ
	MGS annotées jusqu'à		
	Domaine	1 (0%)	3 (0,1%)
	Classe	3 (0,1%)	3 (0,1%)
	Famille	0 (0%)	1 (0%)
	Genre	1 960 (74,6%)	1 988 (74,5%)
	Espèce	665 (25,3%)	673 (25,2%)

	MetaChick spécifique	**Intersection**	**Enrichi spécifique**
Nombre de taxons			
Domaine	0	2	0
Phylum	0	24	0
Classe	0	31	0
Ordre	0	62	2
Famille	0	137	2
Genre	0	649	7
Espèce	1	664	8

KO : KEGG Orthologue, *i.e.* fonctions différentes KEGG ; pathways : voies métaboliques ou catégories fonctionnelles KEGG ; MQ et HQ : qualité moyenne et haute (en fonction de la complétude et la contamination.

E.2.3. Métagénomique quantitative fonctionnelle et taxonomique : méthodologie et filtres

L'étape suivant la constitution d'une référence fonctionnelle et taxonomique est la métagénomique quantitative, autrement dit la quantification des gènes et des MGS (**Figure E-2**).

Quantification fonctionnelle

Comme indiqué en introduction, la quantification des gènes est relativement simple. Elle consiste au réalignement des lectures sur les séquences des gènes, puis au calcul de l'abondance des gènes (exprimée en nombre de copie) en fonction du nombre de lectures alignées.

Les lectures de nos 39 échantillons s'alignent en moyenne à 83.1% ± 0,4% sur 5,1 millions de gènes du catalogue enrichi. Toutefois, sur l'ensemble de la table de comptage, 65,3% des abondances sont nulles, et 82,6% sont inférieures à 1, ce qui illustre la grande parcimonie* et la proportion d'abondance très faible de cette table de comptage.

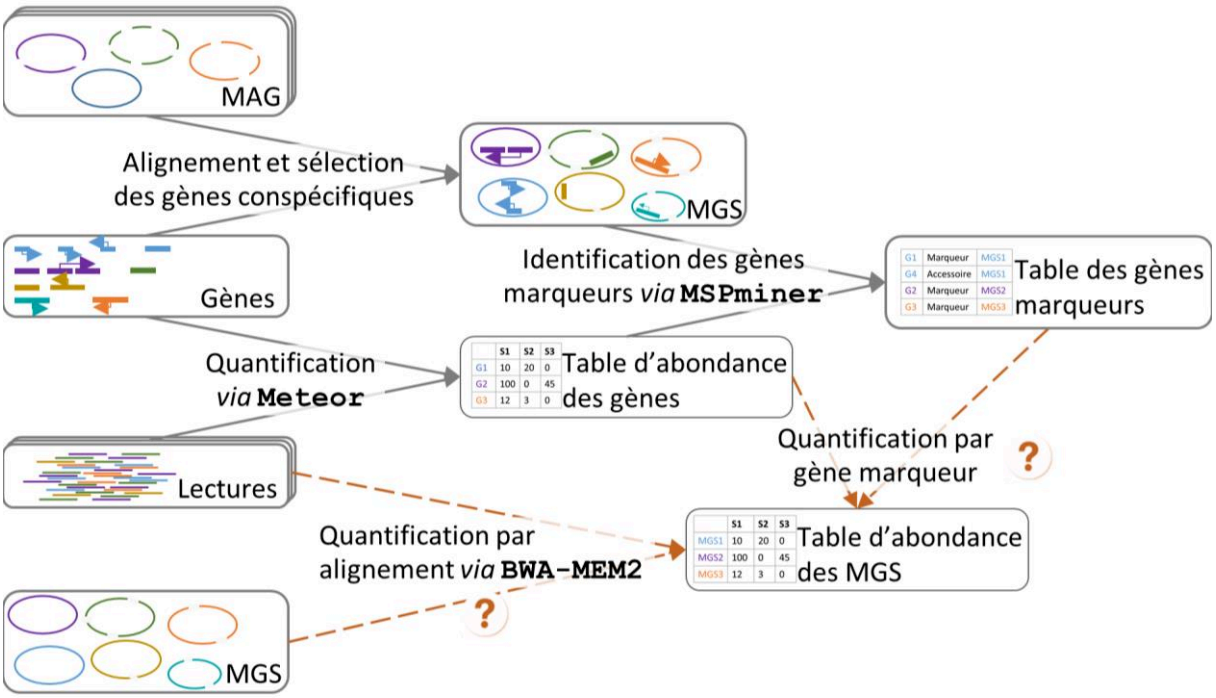


Figure E-2 : Schéma du processus de traitement de métagénomique quantitative.
Alors que la quantification des gènes est relativement standardisée, celle des MGS fait appel à différentes stratégies (flèches en pointillées).

* Parcimonie : proportion de valeur à 0 dans un tableau.

Quantification taxonomique

En ce qui concerne la quantification des MGS, deux types de méthodes existent et nous avons voulu les comparer (**Figure E-2**). La première est équivalente à celle utilisée pour les gènes, *i.e.* réalignement des lectures sur les MGS pour calculer la profondeur de couverture, autrement dit un nombre de copie (méthode référencée par « réalignement » par la suite). La seconde utilise l'abondance moyenne de gènes marqueurs de chaque MGS, également exprimée en nombre de copie (méthode et outils détaillés au chapitre **C.3.2**).

L'application de ces deux méthodes sur notre design expérimental a montré des résultats très différents. Sur les 2 668 MGS du catalogue enrichi, seulement 356 ont pu être quantifiées grâce à la définition de gènes marqueurs, alors que la quantification par réalignement des séquences permet de quantifier 2 415 MGS. En supposant que la diversité de nos écosystèmes cæcaux ait été globalement bien représentée dans notre analyse de métabarcoding (601 ASV), ces résultats suggèrent que l'application de la méthode par gène marqueur sur nos échantillons sous-estime la richesse taxonomique, alors que la méthode par réalignement semble la surestimer.

Pour mieux comparer ces deux stratégies, nous les avons appliquées sur les MGS du catalogue MetaChick uniquement. En effet, le catalogue MetaChick fournit pour chaque MGS un ensemble de gènes marqueurs qui ont été identifiés grâce à un plus grand nombre d'échantillons, séquencés de façon homogène à forte profondeur. De cette façon, nous éliminons le biais de notre échantillonnage (peu d'échantillons avec des profondeurs de séquençage variables) qui pourrait expliquer le peu de MGS pour lesquelles nous réussissons à identifier des gènes marqueurs.

Sur les 2 629 MGS du catalogue MetaChick, nos échantillons permettent de quantifier 1 094 MGS (41,6%) grâce aux gènes marqueurs et 2 306 MGS (87,7%) par réalignement, dont l'ensemble des MGS quantifiées par gènes marqueur sauf une (**Tableau E-6**). Une grande différence dans la richesse des MGS quantifiées persiste donc entre les deux méthodes.

Tableau E-6 : Comparaison des méthodes de quantification taxonomique appliquées au catalogue MetaChick.

	Quantification par **gène marqueur**	Quantification par **réalignement** des lectures	
		Sans filtre	Avec une couverture minimale >= 10%
Quantifiée dans au moins 1 échantillons	1 094 MGS	2 306 MGS	1 067 MGS
Abondance individuelle = 0	52,1%	20,6%	50%
Abondance individuelle < 1	79,8%	89,7%	79,1%

L'analyse des abondances montre par ailleurs que la parcimonie est plus importante pour la quantification par gène marqueur (52,1% versus 20,6%), mais la proportion des abondances inférieures à 1 est plus importante pour la quantification par réalignement (89,7% versus 79,8%). Ainsi, bien que cette dernière quantification permette de détecter plus de MGS, ces dernières sont présentes en très faible abondance (**Figure E-3-A**).

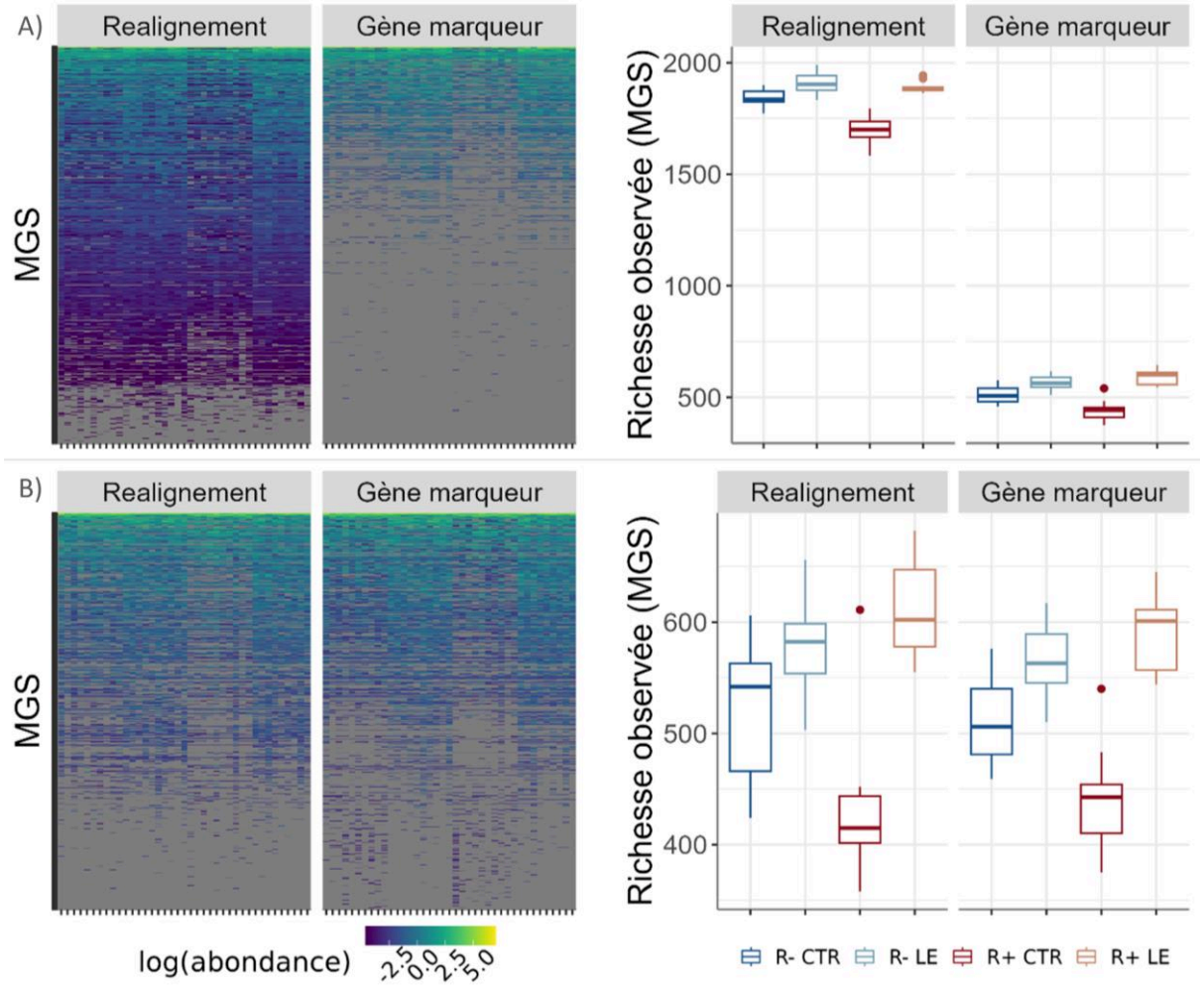


Figure E-3 : Distribution des abondances et richesse des MGS du catalogue MetaChick selon deux méthodes de quantification.
Les quantifications par réalignement sont calculées A) sans filtre ; B) en considérant comme nulle les abondances des MGS couvertes sur moins de 10% de leur séquence.

Une hypothèse pour expliquer cette différence de richesse est la prise en compte par la méthode de réalignement, de MGS dont la couverture* est très faible. Théoriquement, plus une MGS est abondante, mieux elle est séquencée et plus sa couverture est importante. Bien que cela soit statistiquement vrai (R = 0,81, *P*-value < 0,001), sur notre jeu de données 601 MGS couvertes sur moins

* Couverture : nombre de bases couvertes au moins une fois par une lecture.

de 10% de leur longueur ont des abondances supérieures à 1 et peuvent atteindre 17,6. Ces MGS sont donc quantifiées avec un nombre de copies parfois important alors que les lectures comptabilisées dans ces fortes quantifications se concentrent seulement sur un petit morceau de leur séquence.

Ainsi, en éliminant les abondances calculées sur des MGS couvertes sur moins de 10% de leur séquence, 1 067 MGS sont quantifiées par réalignement, dont 909 le sont également avec la quantification par gènes marqueurs, et les profils d'abondance et de richesse des deux méthodes deviennent similaires (**Tableau E-6** et **Figure E-3-B**).

Puisque nous ne pouvons pas améliorer l'utilisation de la méthode par gène marqueur sur notre catalogue enrichi et avec nos échantillons, nous avons utilisé la méthode de quantification par réalignement des lectures en appliquant ce nouveau paramétrage de couverture minimale de 10%. Ainsi, sur notre catalogue enrichi, nous quantifions 1 105 MGS dans au moins un échantillon. Cette table d'abondance présente toutefois une parcimonie encore importante avec 50,2% des abondances nulles et 73,7% des abondances inférieures à 1.

Filtre des gènes et MGS rares

Que ce soit à l'échelle des gènes ou bien à l'échelle des MGS, ces analyses de métagénomique quantitative nous ont montré la proportion très importante d'abondances très faibles (inférieures à 1). Ces gènes et MGS aux abondances et/ou prévalence[*] faibles peuvent être le fruit d'espèces microbiennes rares, mais également celui de bruits de fond dus à la procédure de comptage *via* des lectures courtes, combinés à l'ajout d'éléments (gènes ou MGS) du catalogue MetaChick qui ne sont en réalité pas présents. Pour éviter de tenir compte de ces potentiels gènes et MGS faux positifs, il est communément admis de filtrer selon un seuil minimal d'abondance dans un nombre minimal d'échantillons. Nous avons choisi de conserver les gènes et les MGS avec une abondance en nombre de copies supérieur à 0,75 copie (soit, pour un gène de 1kb, 6 lectures de 145 pb) dans au moins 10% des échantillons (soit 4 échantillons) (**Tableau E-7**).

[*] Prévalence : nombre d'échantillons dans lesquels l'élément est présent

Tableau E-7 : Nombre de gènes et de MGS du catalogue enrichi présents, quantifiés, filtrés.

		Gènes	MGS
Nombre inclu dans le catalogue enrichi		13 765 236	2 668
Nombre quantifié par au moins 1 lecture		5 082 210	1 105
Nombre après filtre sur abondance et prévalence		2 424 181	693
Proportion conservée par échantillons	min	74,4%	75,3%
	moyenne	81,6%	87,9%
	max	86,3%	94,8%
Proportion des abondances conservées par échantillon	min	94,0%	94,2%
	moyenne	98,0%	98,6%
	max	99,4%	99,8%

Pour les gènes, 52,3% des gènes quantifiés sont éliminés, mais ils représentent pour la grande majorité des gènes présents en très faible abondance et dans un très petit nombre d'échantillons. En effet, dans un échantillon au moins 74,4% des gènes quantifiés sont conservés et ils représentent au moins 94% des abondances cumulées dans cet échantillon. Les mêmes tendances sont observées pour les MGS, avec 693 MGS (62,7%) quantifiés après l'application des filtres qui correspondent au sein de chaque échantillon à plus de 75% des MGS détectés et 94% des abondances cumulées.

Pour conclure sur la contribution du catalogue MetaChick et des analyses *de novo* par co-assemblage, 165 MGS sur 693 (24%) sont représentées par des MGS issues de l'analyse *de novo* suggérant une amélioration de la qualité de la séquence représentative. Chacune des analyses a également contribué à enrichir la diversité des MGS détectées sur nos échantillons. Cent trente-trois MGS ont été ajoutées grâce au catalogue MetaChick (dont 58 de haute qualité) et 29 (dont 16 grâce aux échantillons de la condition LE) ont été ajoutées grâce aux co-assemblages (dont 6 de haute qualité). Ces nouvelles MGS spécifiquement apportées par l'une ou l'autre des analyses permettent d'ajouter 12 genres (dont 8 *via* le catalogue MetaChick) et 50 espèces (dont 43 *via* le catalogue MetaChick). Bien que les co-assemblages permettent de quantifier 4 genres et 7 espèces non identifiés sur le catalogue MetaChick, ces genres (à l'exception du genre *JAAWCD01*) et ces espèces ont été détectés sur d'autres échantillons de contenus intestinaux de poules (**Tableau E-8**), ce qui nous conforte dans l'idée qu'ils ne sont pas artéfactuels.

Tableau E-8 : Organismes hôtes des génomes appartenant aux genres et espèces détectés uniquement par co-assemblage.

Genre*	Espèce	Organismes hôtes**	Référence
Dwaynesavagella	*Dwaynesavagella gallinarum*	poule, dinde	(Glendinning et al. 2020; Gilroy et al. 2021; Segura-Wang et al. 2021)
Massilistercora	*Massilistercora timonensis*	humain, poule	(Gilroy et al. 2021)
Rubneribacter	*Rubneribacter badeniensis*	poule, humain	(Glendinning et al. 2020; Gilroy et al. 2021)
Blautia_A	*Blautia_A excrementipullorum*	poule	(Glendinning et al. 2020; Gilroy et al. 2021)
Mediterraneibacter	*Mediterraneibacter intestinigallinarum*	poule, humain	(Gilroy et al. 2021)
Roslinia	*Roslinia cæcavium*	poule	(Glendinning et al. 2020; Gilroy et al. 2021)
Mediterraneibacter	*Mediterraneibacter quadrami*	poule	(Gilroy et al. 2021)
Nanosyncoccus		ruminants variés, souris, homme, porc, éléphant, poule	
JAAWCD01		souris	

* Les genres soulignés sont les 4 genres détectés uniquement *via* les co-assemblages. Deux d'entre eux sont associés à une espèce également spécifique des co-assemblages.
** Pour chaque espèce ou genre, liste des organismes hôtes desquels au moins un génome a pu être assemblé et pris en compte dans la base GTDB r220. Les organismes sont triés par ordre décroissant selon la fréquence d'apparition de leur genre ou de leur espèce parmi les génomes assemblés publiés.

E.3. Effet de la lignée et du régime alimentaire sur la composition microbienne et fonctionnelle : comparaison de la métagénomique et du métabarcoding

E.3.1. La métagénomique améliore la résolution taxonomique mais les affiliations taxonomiques divergent partiellement entre analyses omiques

Les microbiotes cæcaux de notre dispositif expérimental sont composés à partir de 693 MGS affiliées à 99,7% au moins jusqu'au genre (correspondant à 267 genres différents) et à 44,7% à l'espèce (correspondant à 310 espèces différentes). En comparaison à notre étude basée sur l'analyse de la région V3-V4 du gène de l'ARNr 16S, la précision de l'assignation taxonomique est nettement améliorée puisque nous avions 60,6% des 601 ASV affiliés au genre et seulement 20,6% affiliés à l'espèce.

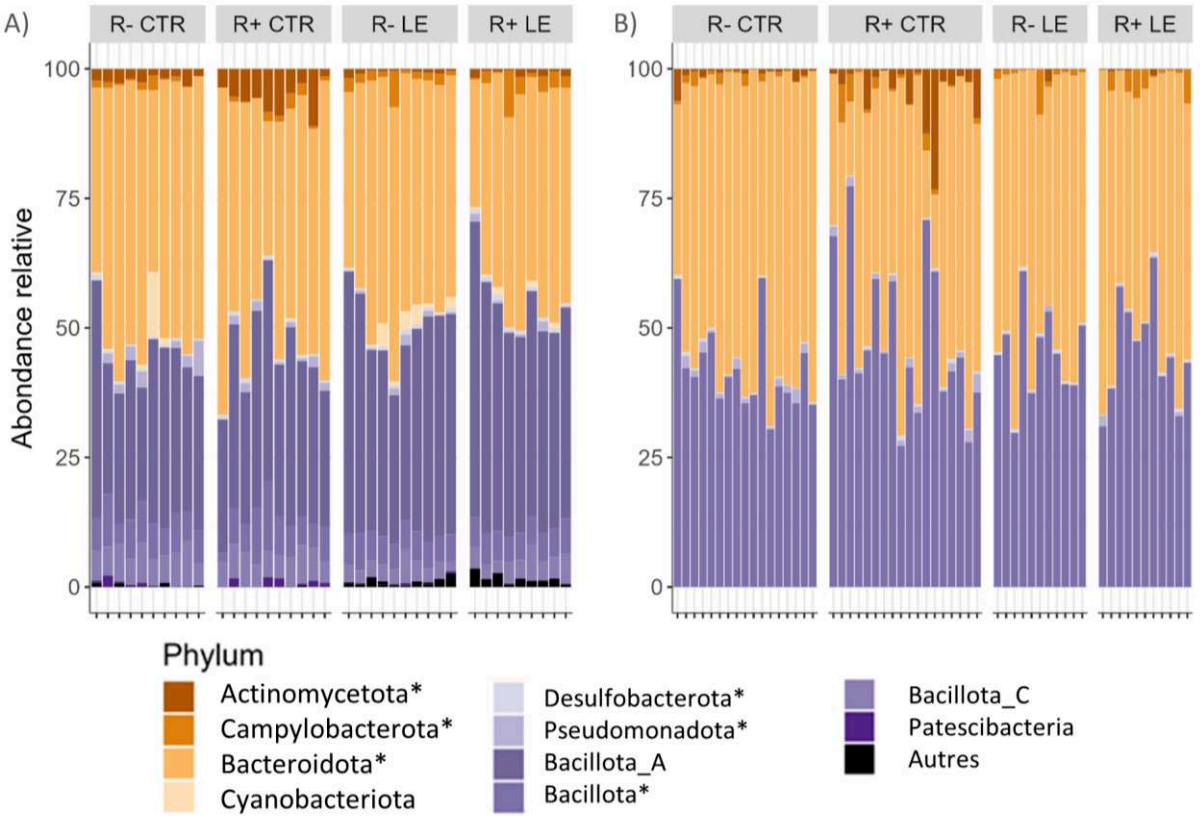


Figure E-4 : Abondance relative des phyla dans chacun des groupes « lignées x régime ».
A) Quantification sur les données de métagénomique ; B) Quantification sur les données de métabarcoding. Les phyla marqués d'une asterisque sont les phyla quantifiés en commun sur les deux omiques.

A l'image des résultats d'analyse de métabarcoding (Tableau 2 de l'article publié dans Scientific Reports, paragraphe **D.5**), en termes de richesse au sein de chaque phylum, les MGS sont dominées par les Bacillota, anciennement Firmicutes (636 MGS contre 500 ASV) puis par les Bacteroidota (22 MGS contre 84 ASV). Toutefois, en termes d'abondance ces deux phyla présentent une répartition presque équilibrée (47.5% pour les Bacillota et 44.8% pour les Bacteroidota sur les MGS contre 45,1% pour les Bacillota et 50,4% pour les Bacteroidota sur les ASV) (**Figure E-4**), ce qui est globalement en accord avec l'âge de nos animaux (31 semaines) (Videnska et al. 2014 et figure **Figure A-6**).

En comparaison avec l'analyse de métabarcoding, l'analyse de métagénomique permet d'identifier de nouveaux phyla, en particulier Thermoplasmatota appartenant au domaine des archées, domaine non détecté par l'analyse de métabarcoding (**Tableau E-9**). Ces différences de détection même à des rangs taxonomiques élevés peuvent s'expliquer par la difficulté en métabarcoding, de définir des amorces universelles permettant d'amplifier l'ensemble des procaryotes présents. C'est en particulier le cas des archées non capturées par les amorces utilisées lors de notre expérience de métabarcoding (Nadkarni et al. 2002).

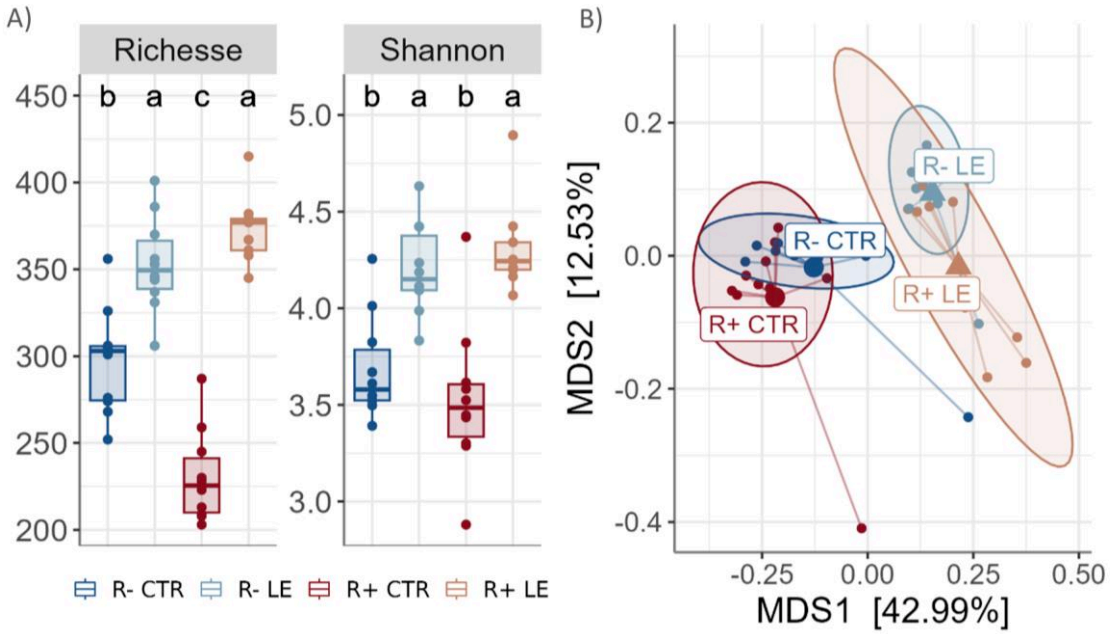


Figure E-5 : Indices de richesse et de diversité des MGS selon la lignée et le régime.
Sur les abondances raréfiées, A) distribution des indices de diversité α; B) analyse en coordonnées principales (PCoA) des distances de Bray Curtis. Les lettres différentes indiquent des différences significatives obtenues par tests post-hoc appliqués à un modèle linéaire prenant en compte la lignée, le régime et leur interaction.

Plus globalement, les analyses de richesse et de diversité confirment les résultats observés en métabarcoding (**Figure E-5**). Les indices d'alpha-diversité sont significativement plus hauts dans la condition du régime LE par rapport à ceux de la condition du régime CTR, confirmant l'effet du régime sur la diversité (*P*-value < 0,001 pour l'indice de Shannon). De plus, les lignées en condition de régime CTR seulement présentent également une différence de richesse, avec la lignée R- plus riche que la lignée R+, confirmant l'interaction du régime et de la lignée sur la richesse (*P*-value< 0,001). Les analyses de dissimilarité bêta confirment également l'effet du régime (*P*-value < 0,001) et un effet de la lignée (*P*-value = 0,03), séparant distinctement les groupes d'échantillons en fonction du régime alimentaire et de façon plus modérée les groupes d'échantillons en fonction de la lignée. Toutefois, ces résultats ne tiennent pas compte des MGS les moins abondants. En effet, pour tenir compte des différences de profondeur de séquençage, ces analyses sont faites sur un nombre de lectures raréfiées (à 52,3 millions de lectures par échantillon). Cette raréfaction impacte de manière importante le nombre de MGS pris en compte (599 parmi les 693 MGS). Au contraire, dans le cadre de l'analyse de métabarcoding, elle n'avait eu aucun effet sur la richesse. L'analyse de l'indice de richesse Chao1 qui estime le nombre de MGS non observées par échantillon, confirme cet effet avec en moyenne 189 MGS non observées par échantillon (contre 38 sur l'analyse de métabarcoding) et jusqu'à 313. Ainsi, la profondeur de séquençage des échantillons de la condition LE (3^{e} phase du projet MetaChick), qui

définit le seuil de raréfaction, n'est peut-être pas suffisante pour représenter l'ensemble des micro-organismes des contenus cæcaux de poules.

Tableau E-9 : Abondance relative et effet de la lignée et du régime à l'échelle des phyla.

Domaine	Phylum	nb MGS	Abondance relative ± erreur standard					Limma *P*-values	
			Globale	R- CTR	R- LE	R+ CTR	R+ LE	L	R
Archaea	**Thermoplasmatota**	1	0,1 ±0	0 ±0	0,1 ±0	0 ±0	0,1 ±0	n.s	***
Bacteria	Actinomycetota	14	2,5 ±0,4	2,2 ±0,2	0,7 ±0,1	5,8 ±0,9	1 ±0,2	*	***
Bacteria	Bacillota	103	6,2 ±0,3	6,8 ±0,5	5,6 ±0,2	6,9 ±1,1	5,6 ±0,2	n.s	***
Bacteria	Bacillota_A	527	36,3 ±1,3	30,9 ±2,2	39,8 ±2,1	32,1 ±1,9	43 ±2,1	n.s	*
Bacteria	Bacillota_B	4	0,2 ±0	0,1 ±0	0,2 ±0	0 ±0	0,3 ±0	n.s	**
Bacteria	Bacillota_C	2	4,8 ±0,3	6,1 ±0,4	3,4 ±0,4	5,6 ±0,4	4,2 ±0,5	n.s	***
Bacteria	Bacteroidota	22	44,8 ±1,3	48,4 ±2,3	44,4 ±1,7	45,9 ±3,4	39,9 ±2,1	n.s	***
Bacteria	Campylobacterota	2	1,7 ±0,3	0,9 ±0,3	2,1 ±0,6	1,1 ±0,3	2,7 ±0,9	n.s	n.s
Bacteria	**Cyanobacteriota**	7	0,9 ±0,4	1,5 ±1,2	1,6 ±0,6	0,1 ±0	0,5 ±0,2	n.s	n.s
Bacteria	Desulfobacterota	3	0,6 ±0	0,5 ±0,1	0,6 ±0,1	0,6 ±0,1	0,8 ±0,1	*	**
Bacteria	**Patescibacteria**	2	0,3 ±0,1	0,4 ±0,2	0,1 ±0,1	0,8 ±0,2	0,1 ±0	n.s	*
Bacteria	Pseudomonadota	3	1,1 ±0,2	2 ±0,6	0,6 ±0,2	1,1 ±0,2	0,6 ±0,2	n.s	**
Bacteria	**Spirochaetota**	1	2E-03 ±1E-03	6E-03 ±3E-03	0 ±0	2E-03 ±2E-03	0 ±0	n.s	**

Les phyla en gras représentent les phyla non détectés par l'analyse métabarcoding de la région V3-V4 du gène de l'ARNr 16S. Les effets de la lignée (L) et du régime (R) ont été testés par un modèle linéaire prenant en compte la lignée et le régime (l'interaction n'étant jamais significative) avec le package R Limma (M. E. Ritchie et al. 2015b) sur les abondances relatives normalisées par CLR (*Centered Log Ratio*): n.s = non significatif ; * < 0,05 ; ** < 0,01 ; *** < 0,001.

En termes de variation des abondances relatives à l'échelle des phyla, l'analyse de métagénomique confirme l'effet de la lignée et du régime sur l'abondance des Actinomycetota, anciennement Actinobacteriota, qui sont plus abondantes dans la condition R+ CTR comparé aux 3 autres groupes (**Figure E-4** et **Tableau E-9**). Au contraire, alors que les autres phyla ne présentaient pas de variation d'abondance sur l'analyse de métabarcoding, ici ils sont presque tous significativement impactés par le régime. La comparaison aux rangs taxonomiques plus petits, comme le genre, est rendue difficile d'une part à cause de différences de détection de chacune des méthodes omiques et de leur capacité à annoter les ASV et MGS jusqu'à ce rang ; d'autre part car chacune des analyses utilise des bases de

références différentes (SILVA pour le métabarcoding et GTDB pour la métagénomique) qui utilisent des classifications taxonomiques différentes. Comme indiqué en introduction (paragraphe **A.5.1.c**), SILVA construit ses propres arbres phylogénétiques de séquences d'ARNr, puis repose sur différentes sources pour nommer chaque taxon. GTDB construit également un arbre phylogénétique à partir de génomes complets puis se base principalement sur la taxonomie du NCBI et y apporte quelques corrections en fonction du positionnement des génomes dans l'arbre. A titre indicatif, les 693 MGS sont affiliées à 268 genres connus (c'est-à-dire non classifiés « unknown genus ») alors que les 601 ASV sont affiliées à 76 genres connus. En tenant compte de la parenté (c'est-à-dire la taxonomie complète du règne au genre), seulement 8 genres sont retrouvés en commun. En corrigeant les noms des phyla de la taxonomie SILVA pour prendre en compte les dernières règles de la nomenclature ICNP (NCBI Staff 2021), 13 genres sont retrouvés en commun, et enfin si on compare seulement les noms de genre sans tenir compte de leur parenté seulement 33 noms sont en communs.

Malgré ce fort biais de taxonomies représentées dans chacun des jeux de données, j'ai cherché à valider les résultats principaux observés sur l'analyse de métabarcoding, en comparant les ASV et MGS différentiellement abondants.

La métagénomique confirme les résultats généraux des analyses d'abondances différentielles appliquées aux ASV (**Figure E-6-A**) : i) l'effet de la lignée est majoritairement observé dans la condition CTR et très modéré dans la condition de régime LE ; ii) l'effet du régime est plus important sur les abondances des MGS de la lignée R+ que sur ceux de la lignée R-. Par ailleurs plusieurs MGS sont détectées différentiellement abondantes à la fois entre lignées (sous la condition de régime CTR) et entre régimes (au sein de la lignée R+). Ces MGS suivent un schéma particulier: elles sont en surabondance à la fois dans la condition de régime LE et dans la lignée efficiente R- comparée à la lignée non efficiente R+ nourrie avec le régime CTR (**Figure E-6-B**). Ceci suggère la particularité du microbiote des poules non efficiente R+ nourries avec le régime CTR, et les points communs entre le microbiote de la lignée efficiente R- et ceux provenant de la condition du régime LE.

A)

Comparaison des lignées		Comparaison des régimes	
Au sein du régime	Au sein du régime	Au sein de la lignée	Au sein de la lignée
LE	CTR	R+	R-
10	10	30	100	74

B)

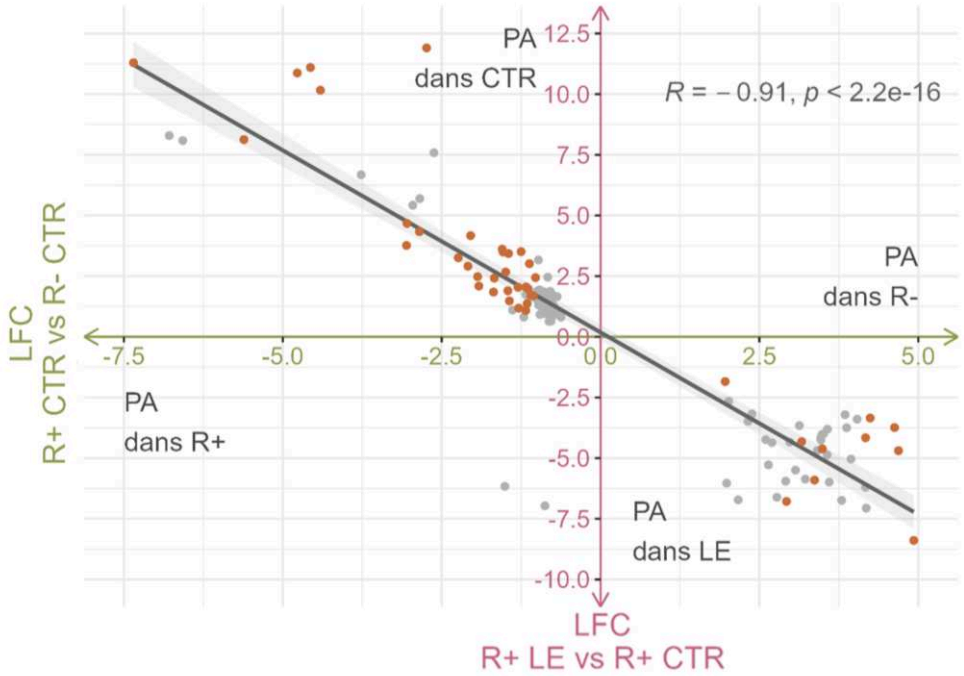


Figure E-6 : Nombre de MGS différentiellement abondant et distribution des log2foldchange (LFC). A) Le tableau indique le nombre de MGS différentiellement abondant (*P*-value ≤ 0,05) entre lignées au sein de chaque régime ou entre régime au sein de chaque lignée avec un LFC ≥ 1 et une abondance relative maximum ≥ 0,1% ; B) Le graphique indique la distribution des LFC des 129 MGS à la fois différentiellement abondant entre lignées au sein du régime CTR et ente régimes au sein de la lignée R+ (en gris ; en orange si LFC ≥ 1 et abondance relative maximum ≥ 0,1%). PA : « plus abondant ».

Toutefois, la comparaison des taxonomies associées à ces effets divergent entre les analyses menées sur les deux omiques.

Parmi les genres mis particulièrement en évidence dans notre analyse de métabarcoding, le genre *Subdoligranulum* est majoritairement renommé *Gemmiger* dans la base GTDB. Ces deux genres présentent toutefois des similitudes en termes de comportement dans nos groupes d'échantillons avec des ASV et MGS plus abondants dans le régime CTR ou le régime LE selon l'entité.

Les genres *Olsenella* et *Anaerosporobacter* sont absents de l'analyse de métagénomique, ils ne sont pas non plus représentés parmi les MGS du catalogue enrichi. Ces genres ont également des synonymes dans la base GTDB (7 pour *Olsenella*) qui peuvent expliquer cette différence de détection,

mais ils soulèvent également l'hypothèse que l'analyse de métagénomique n'ait pu reconstruire de MGS de suffisamment bonne qualité, faute de profondeur de séquençage suffisante.

En cohérence avec sa meilleure résolution taxonomique, la métagénomique identifie, pour 5 genres (*Ligilactobacillus, Limosilactobacillus, Fusicatenibacter, Faecalibacterium, Fournierella*) des MGS affiliées à l'espèce qui suivent des profils d'abondance similaires aux ASV détectés en métabarcoding mais dont les espèces sont inconnues. *A contrario*, pour les genres du phylum Bacteroidota (*Bacteroides, Alistipes,* et *Odoribacter*), aucune MGS n'est similaire aux profils d'abondances des ASV de ces mêmes genres. Plusieurs hypothèses peuvent expliquer ces différences de résultats: un défaut discriminatoire et quantitatif du métabarcoding, et/ou pour la métagénomique un défaut de représentativité de certaines espèces dans les MGS (il y a d'ailleurs moins de MGS que d'ASV affiliés au phylum Bacteroidota), et enfin une différence de détection des ASV/MGS différentiellement abondants en raison des différences de méthodes statistiques utilisées.

Pour permettre une comparaison plus exhaustive de ces deux techniques omiques, il nous faudra d'abord placer les deux tables d'abondance taxonomique dans le même référentiel taxonomique. Cela peut se réaliser en affiliant les ASV sur une base de séquences d'ARNr 16S extraites de la base RefSeq (basée sur les taxonomies du NCBI), et pour les MGS d'utiliser les taxonomies synonymes du NCBI disponibles dans la base GTDB.

E.3.2. Les fonctions du microbiote sont globalement conservées malgré une diversité taxonomique variable

Par rapport à l'analyse de métabarcoding, la métagénomique présente l'avantage d'être une mesure directe des gènes et des fonctions portées par le microbiote. Nous avons choisi d'explorer les résultats de deux sources d'annotations correspondant à deux niveaux de précision d'annotation fonctionnelle. L'une est issue des fonctions KEGG, elle permet une annotation globale des gènes, associée à des catégories fonctionnelles hiérarchisées variées, notamment de différentes voies métaboliques ou processus biologiques. La seconde est issue des annotations CAZy permettant d'étudier les CAZymes, enzymes dédiées au métabolisme des carbohydrates (*« Carbohydrate Active enZymes »*), dont la dégradation a notamment lieu lors de la fermentation des fibres et qui donne ensuite lieu à la production d'acides gras à chaine courte (SCFA).

Sur les 13,7 millions de gènes du catalogue enrichi, 9,4 millions ont un gène orthologue issu de la base de données eggNOG, 7,6 millions (55,6%) le sont avec un score (pourcentage d'identité multiplié par

le pourcentage de couverture) supérieur à 30% (**Tableau E-10**). Ce score nous permet d'assurer un niveau de similarité entre un gène et son orthologue issu de la base de données eggNOG. Les deux séquences doivent ainsi partager 100% d'identité sur au moins 30% de la longueur du gène, ou au moins 30% d'identité sur 100% de la longueur du gène. Parmi ces gènes annotés, 3,9 millions possèdent une annotation avec l'une des deux bases de références fonctionnelles, soit 28,3% de l'ensemble des gènes. Les analyses fonctionnelles des microbiotes cæcaux porteront donc sur une petite partie seulement de l'ensemble des gènes identifiés mais ces taux d'annotation sont globalement pertinents avec ceux d'autres études chez l'homme, le porc ou les ruminants (Almeida et al. 2021; Chen et al. 2021; Xie et al. 2021). En comparaison chez la poule, le catalogue GG-IGC (Feng et al. 2021), qui inclut 16,6 millions de gènes, les taux d'annotation sont proches pour eggNOG (68,2%) et KEGG (30,1% pour 10 665 fonctions KEGG versus 9 336 dans notre étude), mais inférieurs pour les CAZymes (3,4% de la totalité des gènes versus 0,8% dans notre étude).

Tableau E-10 : Nombre et pourcentage de gènes (quantifiés) annotés avec KEGG et/ou CAZy.

	Catalogue enrichi	**Gènes quantifiés**
Nombre de gènes	13 765 236	2 424 181
Nombre de gènes avec un orthologue eggNOG	9 428 135 (68,49%)	1 888 304 (77,89%)
Nombre/% de gènes annotés KEGG/CAZy	4 298 329 (31,2%)	934 004 (38,5%)
Nombre/% de gènes annotés KEGG/CAZy (score > 30)	3 891 468 (28,3%)	875 900 (36,1%)
Annotés avec KEGG	3 891 468 (100%)	875 900 (100%)
Annotés avec CAZy	95 978 (2,4%)	19 748 (2,3%)

Les gènes quantifiés respectent les filtres d'une abondance de 0,75 dans au moins 10% des échantillons. Le score correspond au pourcentage d'identité x le pourcentage de couverture / 100 entre un gène et son orthologue eggNOG.

Parmi les 2.4 millions de gènes quantifiés sur nos échantillons, 0,9 million sont annotés avec KEGG et permettent d'identifier 6 089 fonctions KEGG et 19 748 sont identifiés comme des CAZymes.

Une majorité des fonctions KEGG (60,6%) sont impliquées dans les différentes voies métaboliques, en particulier celles des carbohydrates, des acides aminés et de l'énergie (**Figure E-7**). Dans une moindre importance, elles sont impliquées dans les processus d'interaction avec l'environnement, cellulaires, et génétiques.

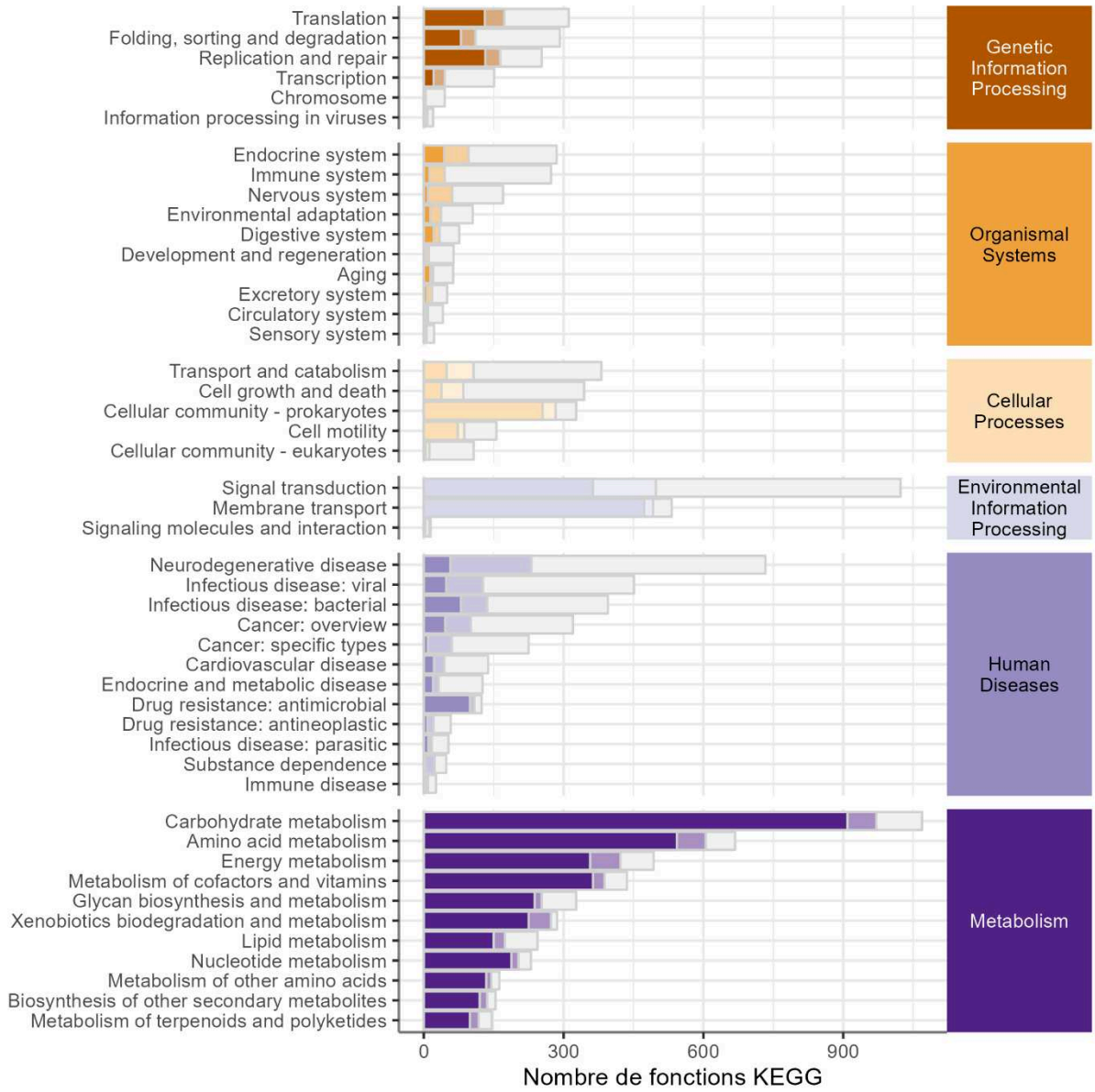


Figure E-7 : Nombre de fonctions KEGG par catégories fonctionnelles.
En gris clair, sur la totalité des gènes du catalogue enrichi, en couleurs claires sur les gènes quantifiés sur nos échantillons, en couleurs foncées sur les gènes quantifiés après filtres sur abondance (> 0,75) et prévalence (dans 10% de nos échantillons).

Il est à noter que sur l'ensemble du catalogue enrichi plus de 22% des fonctions sont incluses dans la catégorie fonctionnelle des maladies humaines, qui représente alors la deuxième catégorie la mieux représentée. La mise en évidence de cette catégorie fonctionnelle est peut-être le reflet d'un biais de représentativité de la base de données KEGG vers les études liées à l'homme et sa santé. Par ailleurs, les fonctions qui sont incluses dans cette catégorie ne sont généralement pas limitées à leur rôle en santé humaine mais impliquées dans d'autres catégories ou métabolismes. Après quantification sur nos échantillons, puis filtres sur l'abondance et la prévalence des gènes, cette catégorie devient

nettement moins représentée (7% des fonctions), à l'exception du sous ensemble de fonctions liées à la résistance antimicrobienne. Cette répartition fonctionnelle paraît plus en cohérence avec notre dispositif expérimental de microbiote cæcal de poule et pointe la présence potentielle de gènes par exemple impliqués dans l'antibiorésistance présents dans nos échantillons et plus largement dans les échantillons du projet MetaChick. Cette capacité de résistance du microbiote est par ailleurs un domaine important de recherche notamment au sein du projet MetaChick (Sidibe et al. 2023).

Comme indiqué précédemment, les CAZymes sont impliquées dans le métabolisme des carbohydrates en permettant notamment la décomposition de structures complexes en molécules plus simples. Elles se répartissent en six classes, elles-mêmes subdivisées en familles de gènes. Les six classes sont : (i) les glycoside hydrolases, la classe la plus représentée avec 1,7 million de séquences dans la base CAZy, permettent la coupure des liaisons glycosidiques tout comme (ii) les polysaccharide lyases (64 896 séquences) ; (iii) les glycosyltransférases, 2ème classe la plus représentée dans la base CAZy avec 1,4 million de séquences, quant à elles permettent la formation des liaisons glycosidiques ; (iv) les estérases de carbohydrate (0,17 million de séquences) interviennent dans l'hydrolyse des esters de carbohydrate ; et (v) les classes des activités auxiliaires (92 735 séquences) et (vi) des modules de fixation des carbohydrates (0,5 million de séquences) facilitent l'activité des autres CAZymes, en incluant des enzymes qui catalysent des réactions d'oxydoréduction ou en permettant la fixation des substrats à proximité des CAZymes. Ces enzymes représentent en général de 1 à 5% des génomes (Lombard et al. 2014) mais sont plus faiblement représentées dans génomes eucaryotes comme ceux de l'homme ou de la poule. A ce jour, la base de données CAZy inclut pour l'assemblage de référence du génome humain 387 gènes de CAZymes soit 0,6% des gènes (et 6 705 séquences protéiques publiées). Elle ne compte pour la poule, que 174 séquences protéiques publiées, l'assemblage de référence de la poule n'ayant pour le moment pas été annoté pour les CAZymes. Ces observations suggèrent que l'annotation eggNOG de nos écosystèmes cæcaux sous-estime probablement la richesse des CAZymes présentes et nécessiterait une annotation spécifique de ce type de protéines.

Dans nos échantillons cæcaux et comme dans de précédentes études sur le microbiote intestinal de poule (Feng et al. 2021; Segura-Wang et al. 2021), les deux classes majoritaires de CAZymes sont les glycoside hydrolases (11 322 gènes répartis dans 51 familles) et les glycosyl-transférases (9 457 gènes répartis dans 28 familles) suivies des modules d'adhésion aux carbohydrates (1 523 gènes répartis dans 7 familles) (**Figure E-8**).

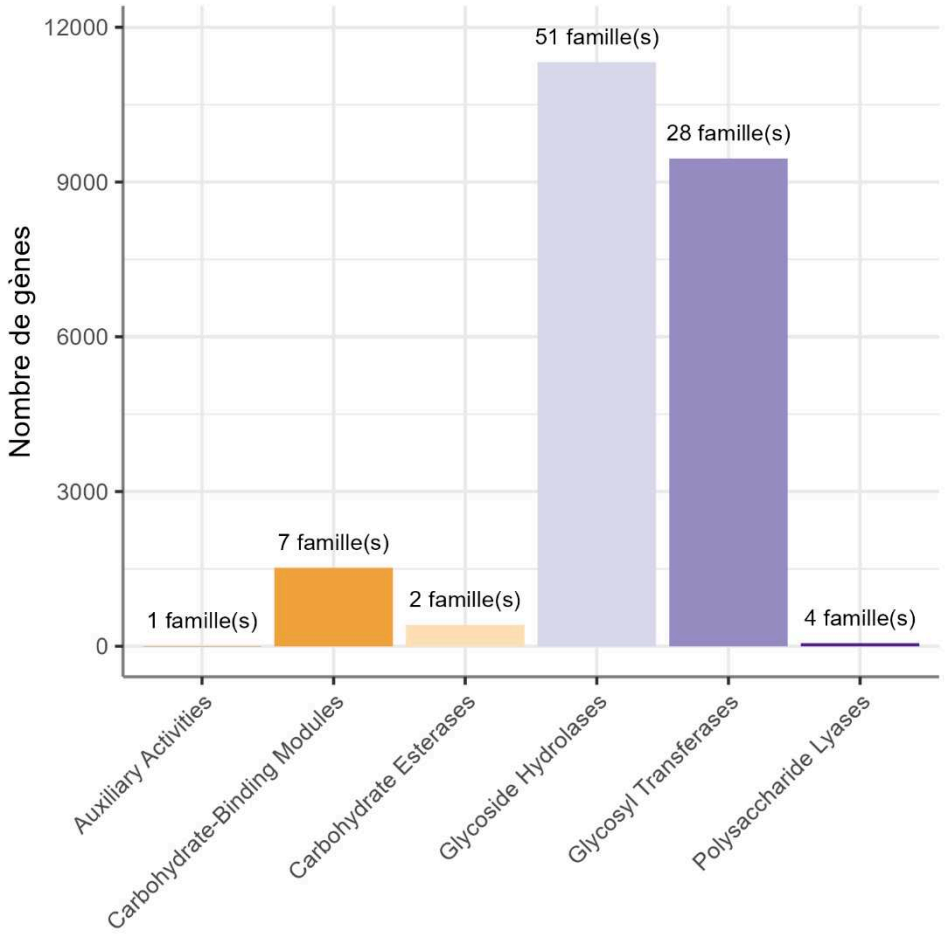


Figure E-8 : Nombre de gènes et de familles de chaque classe de CAZymes, quantifiés sur nos échantillons cæcaux de poule.

Tout comme à l'échelle des MGS, les analyses de richesse et de diversité peuvent s'appliquer aux fonctions. Avec des différences importantes de richesse et de diversité observées entre lignées et entre régimes à l'échelle des MGS (**Figure E-5**), nous pourrions nous attendre également à des différences à l'échelle des fonctions. La **Figure E-9** (A et C) indique que pour chacune des deux bases d'annotation, un effet de la lignée est détecté sur la diversité de Shannon (*P*-value = 0,03 pour les fonctions KEGG, *P*-value = 0,02 pour les familles de CAZymes), mais avec des tendances opposées selon l'annotation. De plus la lignée a un effet sur la richesse mais uniquement des familles de CAZymes (*P*-value < 0,01 versus *P*-value- = 0,053 pour KEGG). La lignée efficiente R- est moins diverse et a tendance à être moins riche fonctionnellement que la lignée R+ pour les fonctions KEGG, ce qui ne suit pas les effets lignées observés à l'échelle des MGS. Elle est en revanche plus riche et plus diverse à l'échelle des familles de CAZymes, ce qui est en accord avec les observations faites sur les MGS. Un effet du régime est également observé (*P*-value = 0,04) sur la richesse des familles de CAZymes uniquement, avec une richesse fonctionnelle plus importante des microbiotes de la condition du régime CTR, ce qui cette fois

ne suit pas l'effet observé sur les MGS. Il faut toutefois modérer ces conclusions, d'une part car l'annotation des CAZymes semble incomplète comme vu précédemment, et d'autre part car les tests post-hoc ne confirment pas totalement ces effets. Ces résultats illustrent un effet plus modéré de la lignée et du régime sur les richesses et diversités fonctionnelles des microbiotes cæcaux comparés aux effets observés sur les MGS.

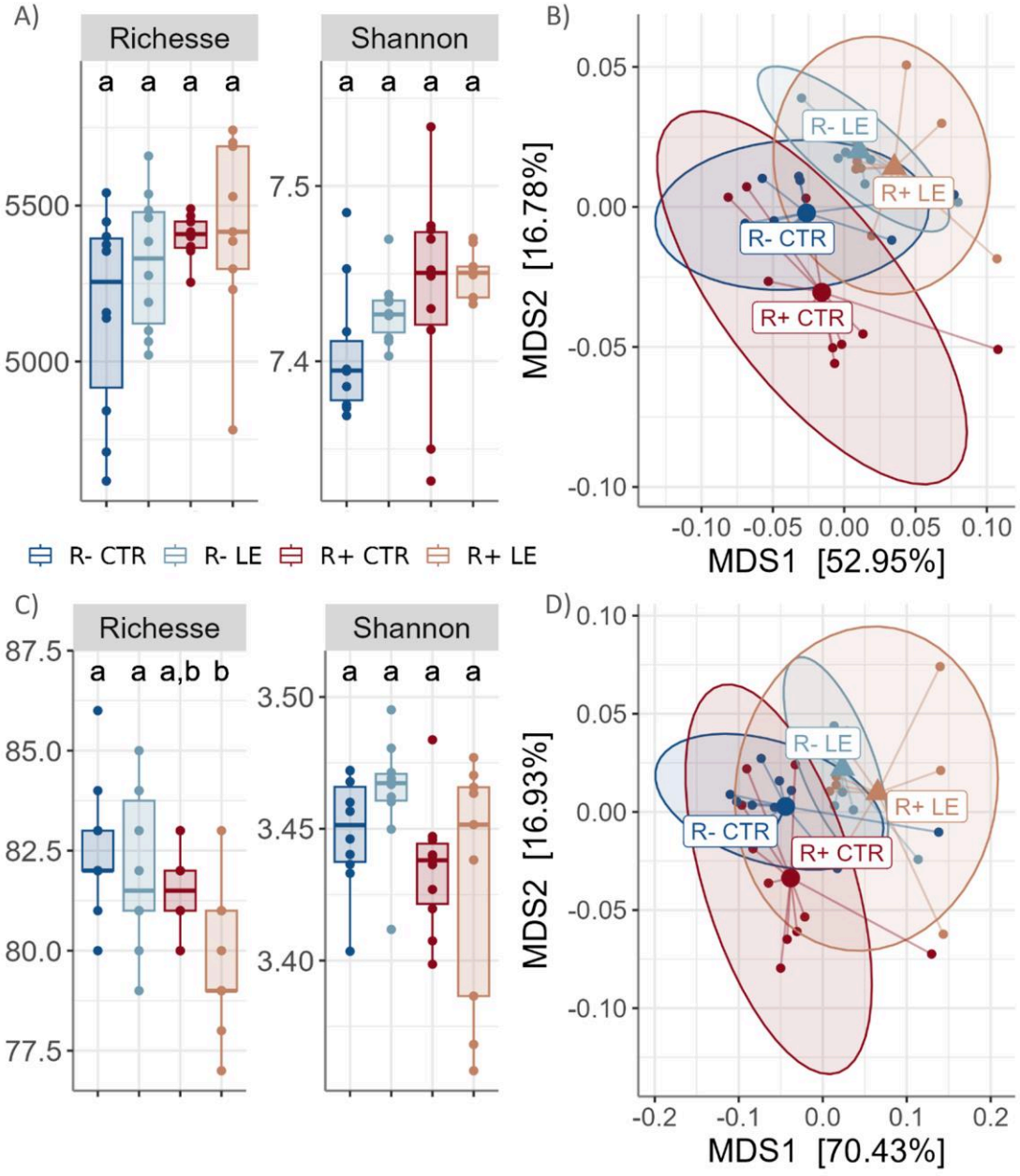


Figure E-9 : Richesse et diversité des fonctions KEGG et des CAZymes selon la lignée et le régime. Sur les abondances raréfiées à gauche (A et C) distribution des indices de diversité α ; à droite (B et D) analyse en coordonnées principales (PCoA) des distances de Bray Curtis, sur les fonctions KEGG en haut (A et B) et les familles de CAZymes en bas (C et D). Les lettres différentes indiquent des différences

significatives obtenues par tests post-hoc appliqués à un modèle linéaire prenant en compte la lignée, le régime et leur interaction.

A l'échelle des dissimilarités entre échantillons, seul l'effet du régime reste significatif (*P*-value < 0.001 pour les fonctions KEGG et les familles de CAZymes). Par ailleurs, en comparaison aux dissimilarités calculées sur les abondances des MGS, les distances entre échantillons calculées sur l'abondance des fonctions sont nettement plus faibles (*P*-values < 0,001) (**Figure E-10**). Tous ces résultats suggèrent une plus grande stabilité du potentiel fonctionnel des microbiotes malgré une grande diversité microbienne et une nette sensibilité des communautés à différents facteurs (lignée et régime). Cette relative stabilité reflète la redondance fonctionnelle des gènes des différents micro-organismes qui colonisent les microbiotes intestinaux. Cette observation a également été montrée chez l'homme sur une étude longitudinale d'une cohorte d'hommes sains (Mehta et al. 2018).

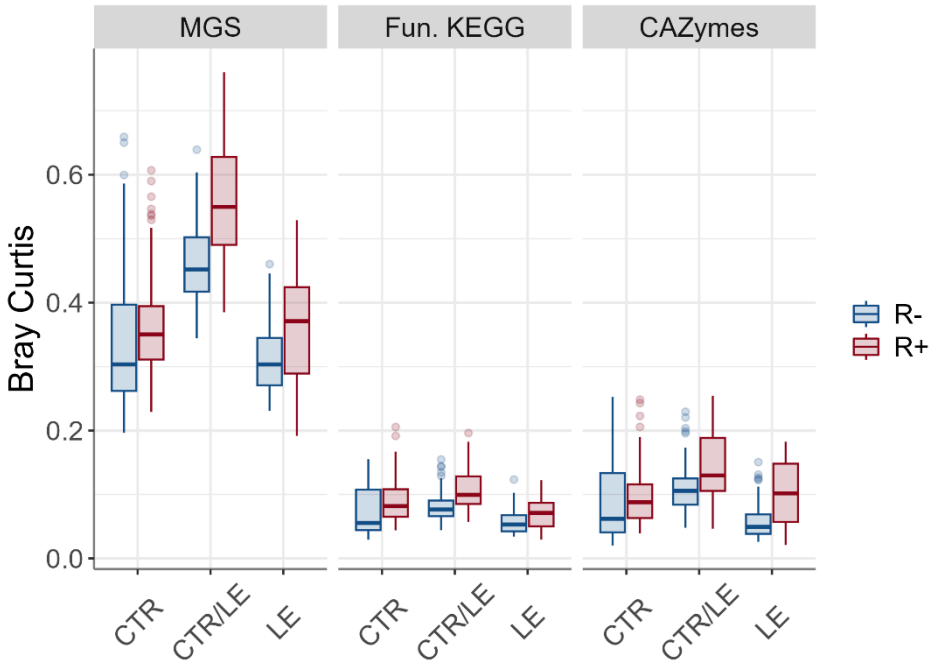


Figure E-10 : Effet du régime sur les distances de Bray Curtis calculées sur les MGS, les fonctions KEGG et les familles de CAZymes potentielles.
Les distances au sein de chaque lignée (R-, R+), calculées sur les abondances des MGS, des fonctions KEGG, des familles de CAZymes, sont prises en compte au sein de chaque régime (CTR, LE) ou entre régimes (CTR/LE).

E.4. Bilan des résultats et réflexion sur le développement méthodologique

Les nombreuses comparaisons ayant permis la mise en place de la procédure de traitement des données de métagénomique illustrent les bénéfices d'une approche *de novo* combinée à la prise en compte d'un catalogue de référence extérieur. L'analyse par co-assemblage a permis de mieux gérer les échantillons séquencés à moindre profondeur, et à contribuer à quantifier de nouvelles espèces non représentées dans le catalogue MetaChick. De son côté, le catalogue MetaChick a permis de compléter l'analyse *de novo* permettant ainsi de couvrir une large diversité taxonomique et fonctionnelle de nos échantillons cæcaux.

Elles ont toutefois mis en évidence des points de faiblesse qui illustrent la non-exhaustivité de ce catalogue enrichi et amènent à penser à de nouveaux développements. A l'échelle de la construction des MGS, une espèce n'a été détectée que sur l'analyse par assemblage individuel. Comme suggéré précédemment, nous pourrions adopter une stratégie hybride d'assemblage en agrégeant les résultats de l'analyse par assemblage individuel, par co-assemblage et en tenant compte du catalogue pré-existant, mais cette étape d'agrégation par déréplication a montré un défaut de pouvoir discriminant de certaines espèces proches amenant également à la perte d'espèce. Pour améliorer cette procédure, nous pourrions imaginer de nouvelles stratégies de sélection des MGS représentant chaque espèce en utilisant à la fois la similarité de séquence (ANI), les annotations taxonomiques, ou encore les abondances et prévalences des gènes :

- Pour différencier les souches, il est communément admis de dérépliquer les MGS à un seuil de 99% d'ANI. A l'image de ce qui est recommandé en métabarcoding pour les stratégies de clustering à seuil fixe, la déréplication pourrait se faire à ce seuil plus élevé. Chaque MGS serait affiliée taxonomiquement puis le meilleur génome représentant chaque espèce serait sélectionné. A noter que cette stratégie peut souffrir du même biais que précédemment si différentes espèces partagent plus de 99% d'ANI. Par ailleurs, pour limiter la consommation de ressources informatiques additionnelles, cette stratégie pourrait être appliquée uniquement aux MGS dont les MAG présentent des valeurs ANI entre 95 et 99%.
- Comme indiqué précédemment, il existe une autre stratégie de définition des espèces qui utilise les informations de prévalence et d'abondance des gènes (marqueurs et accessoires) pour définir des pangénomes d'espèces métagénomiques (MSP). A l'image de ce qui a été réalisé pour la constitution du catalogue MetaChick (Plaza Oñate et al. 2023), des MSP pourraient être générés

pour chaque MGS. Dans le cas de la définition de plusieurs MSP pour une MGS, l'exploitation des profils d'abondances des gènes (notamment accessoires) devrait permettre de différencier les MAG appartenant à différentes espèces[*]. Une MGS par groupe de MAG pourrait enfin être générée. L'intérêt de cette stratégie est qu'elle permet également de faire le lien entre les gènes (support de l'annotation fonctionnelle) et les génomes (support de l'annotation taxonomique). Cependant elle nécessite une profondeur de séquençage et un nombre d'échantillons suffisants.

L'annotation fonctionnelle de son côté semble très partielle. Cette partialité peut être le fruit de fausses détections de gènes, mais il est également reconnu qu'une forte proportion de gènes et de protéines est détectée sans qu'on leur ait attribué de fonction (Berg et al. 2020). Sans que cela soit suffisant, l'utilisation d'outils additionnels potentiellement dédiés à certains types de fonctions ou bases de données pourraient permettre d'améliorer la complétude de cette annotation (par exemple KofamScan pour interroger directement la base de données KEGG (Aramaki et al. 2020)).

En termes d'effet de la lignée et du régime sur la composition taxonomique du microbiote, les données de métagénomique ont confirmé les tendances observées grâce aux analyses de séquences de métabarcoding, mais les analyses comparatives détaillées des taxonomies impliquées dans ces effets font face à la difficulté de ne pas avoir de référentiel commun. Des analyses complémentaires seront menées pour permettre une comparaison exhaustive des taxonomies et abondances associées identifiées dans ces deux omiques (paragraphe **G.2**).

Fonctionnellement, le microbiote cæcal inclut un large panel de fonctions et, bien que la composition taxonomique soit significativement impactée par la lignée et le régime alimentaire, ses fonctions présentent globalement une sensibilité nettement moindre à ces facteurs génétiques et environnementaux. Alors que le régime semble avoir un impact principalement sur les abondances (dissimilarité de Bray Curtis), la lignée semble avoir également des effets sur les abondances et la richesse probablement rare de catégories fonctionnelles plus spécialisées comme les CAZymes.

[*] Un MAG est un ensemble de contigs, une MGS est le MAG représentatif d'un regroupement de MAG partageant (généralement) 95% ANI. Un MSP est un regroupement de gènes co-abondant qui peuvent être préalablement alignés sur les MAG.

F. Partie 3 : Analyse de données de métatranscriptomique : différenciation fonctionnelle des microbiotes cæcaux

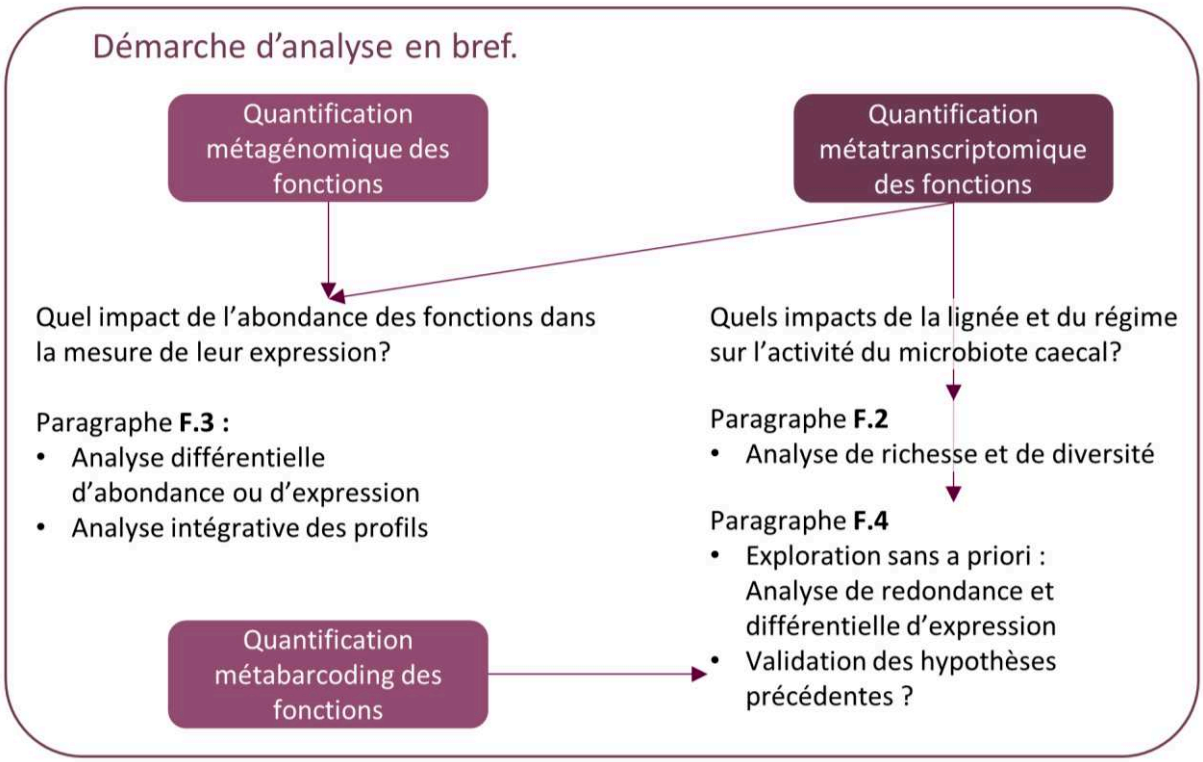
F.1. Contexte et objectifs

Les analyses sur le microbiote intestinal sont majoritairement dominées par l'utilisation de données de séquences de métabarcoding et de métagénomique. Celles-ci permettent en premier lieu de décrire la composition des communautés microbiennes présentes, ainsi que les fonctions portées par ces micro-organismes (de manière directe et théoriquement exhaustive pour la métagénomique).

Dans notre étude, les analyses *via* ces deux techniques omiques nous ont permis d'identifier des effets importants de la lignée et/ou du régime alimentaire sur la composition taxonomique des microbiotes cæcaux. L'analyse de données de métagénomique nous a montré qu'à l'échelle des fonctions ces facteurs impactaient de façon bien plus modérée les microbiotes, illustrant une conservation du potentiel fonctionnel malgré des changements importants de l'environnement (lignée génétique et régime alimentaire). De précédentes études ont aussi observé cette redondance fonctionnelle illustrée par une plus grande stabilité des fonctions que des génomes, chez l'homme (Franzosa et al. 2014; Heintz-Buschart et al. 2016; Mehta et al. 2018) ou encore le bovin (F. Li et al. 2019). Ces études ont également comparé les profils d'abondance taxonomique et fonctionnelle (données de métagénomique) et d'expression fonctionnelle (données de métatranscriptomique). Elles ont montré que les profils d'expression sont plus stables que les profils d'abondance des communautés microbiennes (taxonomiques), mais plus variables que les profils d'abondance fonctionnelle. Par ailleurs, elles ont montré que l'abondance des gènes a un impact sur leur expression. Ainsi, il est probable que les familles de gènes différentiellement exprimées soient également dans des abondances différentes. Toutefois, il existe une part de la variabilité de l'expression qui n'est pas expliquée par l'abondance initiale des gènes. En effet, une augmentation ou une diminution de l'expression d'un gène peut être due à une modulation dans l'abondance du micro-organisme porteur de ce gène et/ou résulter de l'activation ou répression de son expression. Il est ainsi possible d'identifier des fonctions portées par des gènes différentiellement transcrits mais non différentiellement abondants et inversement. Les résultats fonctionnels obtenus sur des données de métabarcoding ou de métagénomique représentent donc les activités potentielles du microbiote mais ne présagent pas systématiquement de ses activités réelles.

L'analyse des données de métabarcoding nous a permis d'émettre quelques hypothèses fonctionnelles reliant la composition du microbiote avec la quantité d'amidon ingérée (différence entre lignées) ou de fibres (différence entre régimes). En particulier, les fonctions des métabolismes de l'amidon et de divers autres carbohydrates semblent différentiellement abondantes entre lignées et entre régime

ainsi que celles du métabolisme du propionate, un acide gras à chaine courte influençant notamment le métabolisme du glucose de la poule. L'analyse des données de métagénomique, quant à elle, nous a permis d'aboutir à la constitution d'une référence de 13,7 millions de gènes dont 2,4 millions sont quantifiés dans nos échantillons (et dont 0,9 million sont annotés). Pour la présente étude, des échantillons du même dispositif expérimental et dont une majorité a également servi à construire cette référence de gènes, ont été séquencés en métatranscriptomique. L'analyse de ce troisième séquençage combinée au profil d'abondance fonctionnelle obtenu sur les données de métagénomique devrait nous permettre: i) d'évaluer la contribution des abondances dans l'expression des fonctions ; ii) d'évaluer la ressemblance fonctionnelle de ces deux omiques ; et iii) d'identifier les fonctions réellement différentiellement actives entre microbiotes selon la lignée de poule et/ou le régime alimentaire afin d'apporter de nouveaux éléments quant à l'association entre le microbiote cæcal, l'efficience alimentaire et l'adaptation à une modification du régime.



Nota Bene : Afin de simplifier la lecture, bien que ce soient les gènes qui soient exprimés et non les fonctions, je ferai référence dans la suite du manuscrit, « d'abondance de fonction » ou de « métagénomes » pour les quantifications des fonctions issues des données de métagénomiques, et « d'expression de fonction » ou de « métatranscriptomes » pour les quantifications issues des données de métatranscriptomique. Par ailleurs, j'ai choisi dans cette analyse de ne pas distinguer une différence

d'expression due à une différence d'abondance, de celle due à une activation/répression de l'activité transcriptionnelle propre à chaque cellule.

F.2. La richesse et la diversité fonctionnelle à l'échelle des profils d'expression

Le séquençage d'ARN messagers (ARNm) dans un contexte d'analyse d'écosystème microbien nécessite au préalable une déplétion des ARN ribosomaux (ARNr) qui représentent la quasi-totalité des ARN dans un échantillon (plus de 97% sur nos échantillons tests non ribodéplétés, paragraphe **C.3.3.a**). Dans cette étude, les ARN de 41 échantillons de contenu cæcal ont été extraits, puis ribodéplétés et séquencés avec en moyenne 43,4 millions de paires de lectures par échantillon. Le contrôle *a posteriori* de la ribodéplétion valide une forte augmentation de la proportion des ARN non ribosomaux, puisqu'ils représentent en moyenne 84% des lectures même si 4 échantillons ont encore entre 31,1 et 41,5% de lectures provenant d'ARNr (**Figure F-1**).

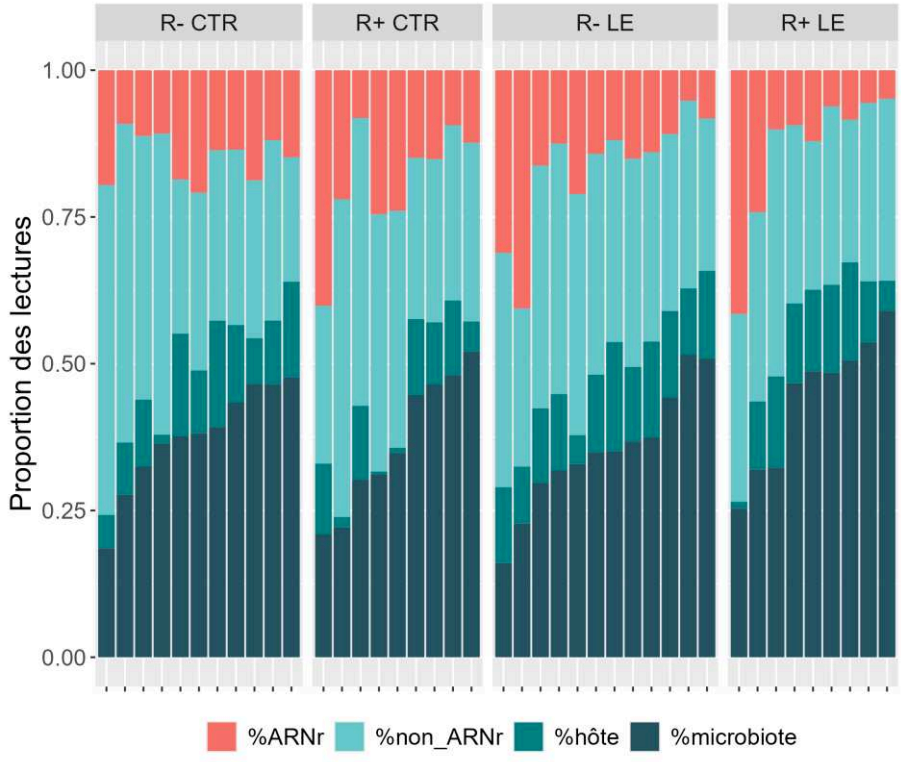


Figure F-1 : Proportion des lectures ribosomales, non ribosomales, provenant de l'hôte ou du microbiote cæcal.
Les lectures non ribosomales provenant de la poule ou du microbiote cæcal correspondent aux lectures non ribosomales alignées sur le génome de la poule ou sur le catalogue enrichi de gènes.
ARNr : ARN ribosomal.

Les lectures non ribosomales sont ensuite nettoyées de leur potentielle contamination par des ARN issus de l'hôte, puis alignées sur le catalogue enrichi de 13,7 millions de gènes. De façon surprenante, le taux d'alignement des lectures sur le catalogue de gènes varie très graduellement de 26,2 à 72,8% selon l'échantillon. Ceci aboutit à un nombre de lectures utiles par échantillon très variables allant de 12,2 millions à 51,8 millions de lectures utiles (soit de 16 à 59% des lectures initiales) avec notamment 5 échantillons ayant moins de 20 millions de lectures alignées. Les lectures qui ne correspondent ni à des ARN ribosomaux ni à des gènes du microbiote cæcal s'expliquent en partie par une contamination par les ARN non ribosomaux de la poule (contamination moyenne à 11%), mais en moyenne 34,8% restent non identifiées. Il pourrait s'agir de séquences provenant de transcrits nouveaux qui ne seraient pas inclus dans le catalogue, mais cette hypothèse semble peu probable pour expliquer la majorité d'entre elles. En effet, le catalogue a été construit à partir de 359 échantillons incluant notamment 35 des 41 échantillons séquencés ici, et tous présentent un taux d'alignement des lectures métagénomiques homogène supérieur à 80%. Sur notre projet pilote, nous avions également observé, entre les deux kits de ribodéplétion, une différence du taux d'alignement sur le catalogue MetaChick des lectures identifiées comme non ribosomales. Ces séquences pourraient donc refléter un biais technique dû au protocole de ribodéplétion avec le kit RiboPOOLs sans pour autant identifier ce que ces séquences représentent. Nous avons choisi de supprimer un échantillon du groupe R- LE conservant le moins de séquences alignées pour ne conserver dans les analyses suivantes seulement des échantillons avec un minimum de 16,7 millions de lectures alignées (moyenne à 32,9 millions).

L'alignement des lectures sur le catalogue enrichi permet d'identifier 3 millions de gènes exprimés représentés par au moins une lecture dans au moins 1 échantillon. Après filtre sur une abondance (exprimée en nombre de transcrits) supérieure à 0,75 dans au moins 10% des échantillons (soit 4 échantillons), 0,81 millions sont conservées et considérés comme réellement exprimés. Ce filtre sur l'abondance et la prévalence* impacte plus fortement les données de métatranscriptomique que les données de métagénomique précédemment. En moyenne 65,3% (minimum à 53,6%) des gènes exprimés dans un échantillons sont conservés (contre 81,6% en métagénomique), mais ils correspondent toutefois à en moyenne 97% (minimum à 94%) de l'expression globale de chaque échantillon. Parmi ces gènes exprimés, 46,4% sont annotés avec une fonction KEGG permettant d'identifier 4 694 fonctions.

* prévalence : nombre d'échantillons dans lequel le gène est présent.

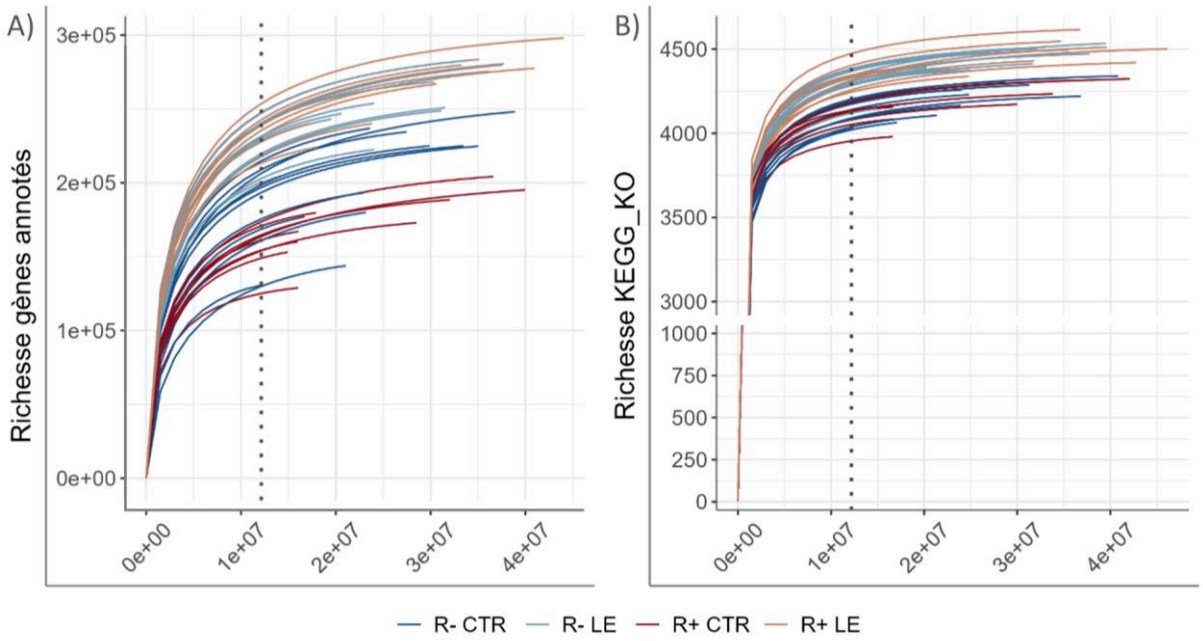


Figure F-2 : Courbes de raréfaction fonctionnelle sur les données de métatranscriptomique.
Courbes calculées à partir des quantifications en nombre de lectures alignées sur les 0,9 millions de gènes filtrés sur leur prévalence et abondance. A) sur la richesse en gènes annotés, B) sur la richesse en fonction KEGG. La ligne en pointillés indique le seuil de raréfaction.

Malgré une variabilité importante du nombre de lectures par échantillon avec finalement un minimum assez faible, les courbes de raréfaction montrent une bonne représentativité de la richesse fonctionnelle des annotations KEGG (**Figure F-2-B**). A l'échelle des gènes annotés pour ces fonctions KEGG, la richesse continue d'augmenter avec un nombre de lectures croissant même si elle augmente moins rapidement pour les échantillons avec le plus de lectures (**Figure F-2-A**). Pour les analyses de richesse, de diversité et de dissimilarité des échantillons, la raréfaction de l'expression des gènes pour tenir compte des différences de nombre de lectures alignées pour chaque échantillon pourrait partiellement impacter la richesse des gènes pris en compte vers ceux les plus abondants, sans pour autant impacter la richesse des fonctions KEGG.

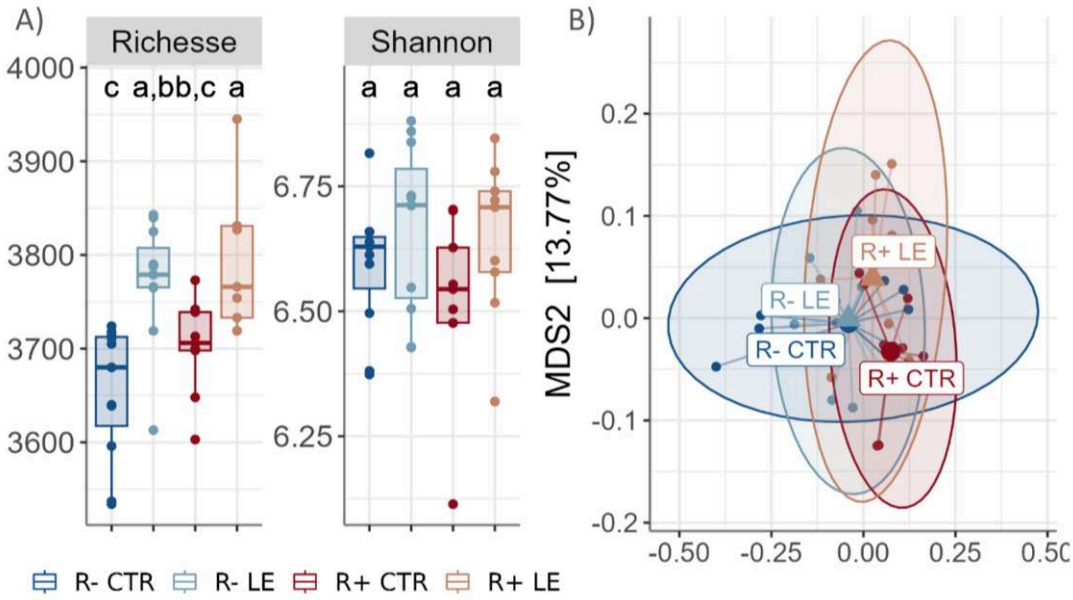


Figure F-3: Richesse et diversité des fonctions KEGG exprimées selon la lignée et le régime.
Sur les expressions raréfiées, A) distribution des indices de diversité α ; B) analyse en coordonnées principales (PCoA) des distances de Bray Curtis. Les lettres différentes indiquent des différences significatives obtenues par tests post-hoc appliqués à un modèle linéaire prenant en compte la lignée, le régime et leur interaction.

Les analyses de richesse et de diversité fonctionnelle montrent une sensibilité des métatranscriptomes à la lignée et au régime. La richesse des fonctions KEGG est différente selon le régime (*P*-value < 0,001), avec les microbiotes du régime LE plus riches en fonctions exprimées KEGG que les microbiotes du régime CTR, mais leurs diversités ne sont impactées ni par le régime ni par la lignée (**Figure F-3-A**). En comparaison aux analyses d'alpha diversité réalisées à l'échelle des espèces métagénomiques (MGS) et des fonctions potentielles KEGG (métagénomes), les métatranscriptomes semblent plus variables que les métagénomes puisque sensibles au régime alimentaire, mais moins variables que les MGS puisque non sensibles à la lignée (**Tableau F-1**).

Tableau F-1 : Comparaison des effets de la lignée et du régime sur les indices de richesse et de β-diversité estimés sur les taxonomies ou les fonctions.

	Taxonomie (MGS)		**Métagénome (KEGG)**		**Métatranscriptome (KEGG)**	
	Richesse	**β diversité**	**Richesse**	**β diversité**	**Richesse**	**β diversité**
Lignée	0,01*	0,03*	0,05	0,05	0,12	0,02*
Régime	< 0,001*	< 0,001*	0,31	< 0,001*	< 0,001*	0,05*
Interaction	< 0,001*	0,06	0,41	0,33	0,48	0,29

Analyse de variance (Anova) à partir d'un modèle linéaire prenant en compte la lignée, le régime et leur interaction. Les astérisques mettent en évidence les effets significatifs (*P*-value < 0,05).

Les analyses de dissimilarité des échantillons en fonction de l'expression des fonctions KEGG (*via* la distance de Bray Curtis), malgré une séparation graphique des groupes d'échantillons peu évidente, confirme l'effet du régime détecté précédemment sur les MGS et sur les métagénomes (*P*-value = 0,048), ainsi que l'effet de la lignée détecté précédemment uniquement sur les MGS (*P*-value = 0,02) (**Figure F-3-B** et **Tableau F-1**). Comme pour la richesse, les métatranscriptomes présentent une plus grande variabilité de fonctions que les métagénomes mais une moins grande variabilité que les MGS (**Figure F-4**). Par ailleurs, la variabilité individuelle des métatranscriptomes est également plus grande que celles des métagénomes.

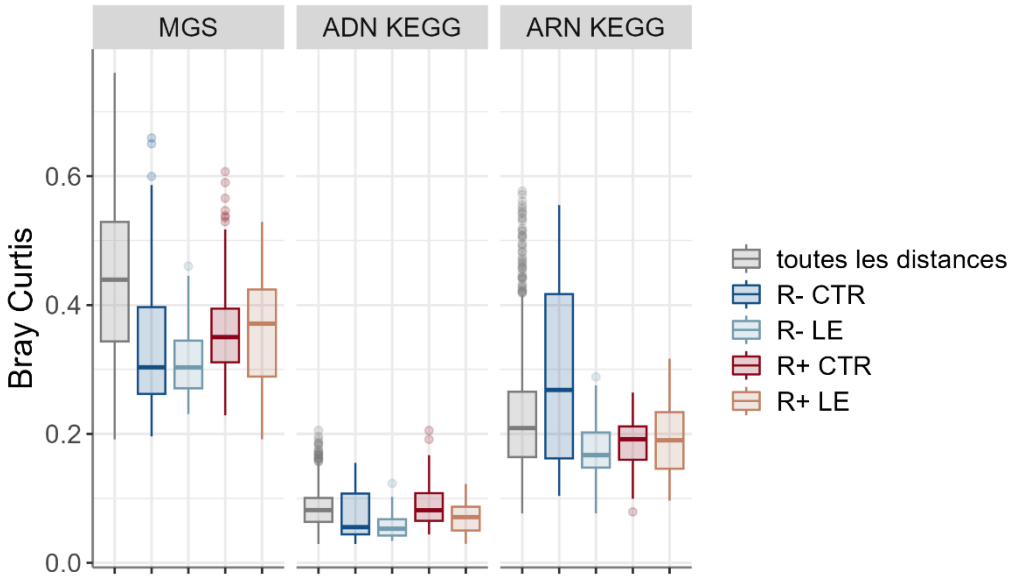


Figure F-4 : Distribution des distances de Bray Curtis calculées sur les MGS, les fonctions KEGG potentielles ou exprimées.
Distribution de toutes les distances entre deux échantillons du même groupe ou de groupe différents (gris) et distributions des distances entre échantillons d'un même groupe « lignée x régime ».

Puisque les métatranscriptomes sont plus variables individuellement et plus sensibles aux facteurs génétique et au régime, nous nous sommes demandées quelle est la part d'une réelle différence d'activité transcriptionnelle du microbiote selon la lignée de l'hôte ou le régime alimentaire, et plus généralement comment quantifier la ressemblance entre les profils métagénomiques et métatranscriptomiques.

F.3. Evaluation de la contribution des profils d'abondance dans les profils d'expression des fonctions

F.3.1. Comparaison des fonctions différentiellement abondantes et/ou exprimées en fonction de la lignée et du régime

Pour identifier précisément les fonctions dont l'abondance ou l'expression est impactée par la lignée ou le régime, j'ai procédé à une analyse différentielle entre lignées au sein de chaque régime, ou entre régime au sein de chaque lignée. Pour cela le package R Limma (M. E. Ritchie et al. 2015a) a été appliqué sur les abondances ou expressions relatives transformées par la méthode CLR sur l'ensemble des fonctions quantifiées (6 089 pour les métagénomes, 4 694 pour les métatranscriptomes).

Tableau F-2 : Nombre de fonctions KEGG différentiellement abondantes et/ou exprimées entre lignées ou entre régimes.

	Comparaisons des lignées		**Comparaisons des régimes**	
	Au sein du régime LE	Au sein du régime CTR	Au sein de la lignée R+	Au sein de la lignée R-
Metagénomes (DA)	37	405	933	603
(dont DA non transcrites)	(26)	(82)	(241)	(152)
Métatranscriptomes (DE)	54	392	1 669	1 368
(dont DE non abondantes)	(0)	(2)	(3)	(6)
DE & DA	0	68	358	242

DA : différentiellement abondante ; DE : différentiellement exprimée. Le seuil de significativité des *P*-value ajustée (BH) est fixé à 0,05.

Le **Tableau F-2** montre que globalement, ces analyses différentielles confirment les tendances observées à l'échelle de la composition des MGS et des ASV. Il y a un fort effet du régime sur les microbiotes provenant deux lignées, et exacerbé sur celui de la lignée non efficiente R+. L'effet de la lignée est plus limité, et majoritairement observé en condition de régime CTR. Elles confirment également les résultats précédents indiquant une plus grande sensibilité des métatranscriptomes que des métagénomes, en particulier à la modification du régime avec 35,5% des fonctions exprimées qui sont détectées variables selon le régime pour les microbiotes de la lignée R+, et 29% pour les microbiotes de la lignée R- (contre 15% et 9,9% pour les métagénomes). Cependant, de façon inattendue, les fonctions différentiellement abondantes (DA) sont relativement peu différentiellement exprimées (DE) et inversement les fonctions différentiellement exprimées sont relativement peu différentiellement abondantes.

Dans le premier cas, cela s'explique partiellement par des fonctions qui ne sont pas du tout exprimées. Ces fonctions quantifiées seulement en métagénomique font globalement partie des fonctions les plus

faiblement abondantes (*P*-value < 0,001) (fonctions représentées par des croix sur la **Figure F-5**). Leur absence dans les données de métatranscriptomique peut refléter un défaut de profondeur de séquençage, ou bien la quantification en métagénomique d'ADN reliques* (Berg et al. 2020).

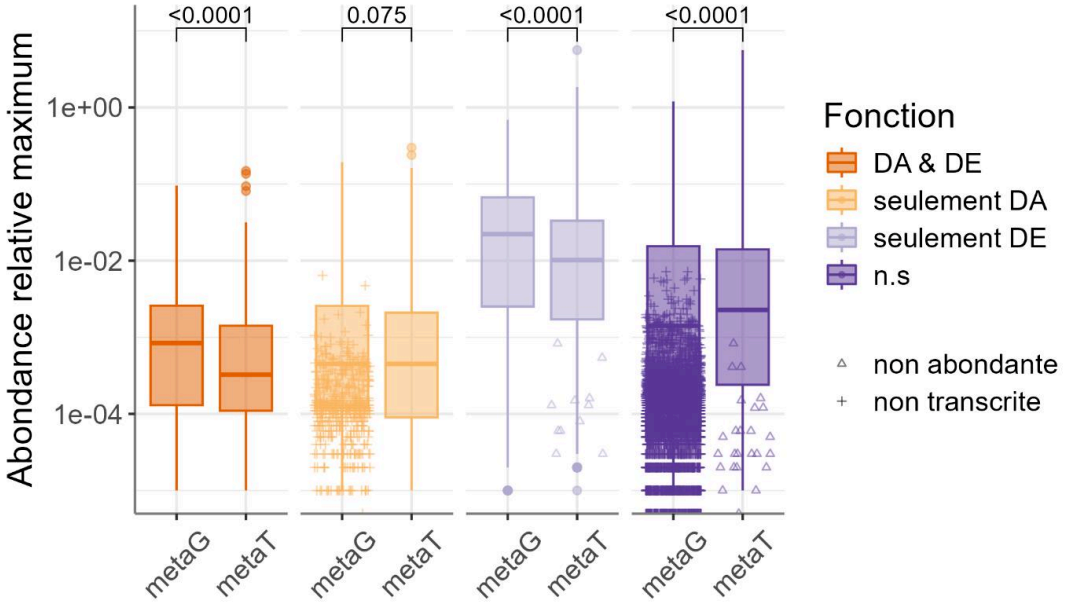


Figure F-5 : Distribution des abondances relatives les plus fortes en métagénomique et en métatranscriptomique.
Pour chaque fonction, chaque comparaison (entre lignées au sein d'un régime ou entre régimes au sein d'une lignée) et chaque omique, sélection de l'abondance/expression relative moyenne la plus forte entre les deux groupes « lignée x régime » comparés. Visualisation selon le statut différentiellement abondante et exprimée (DA & DE), uniquement différentiellement abondante (DA), uniquement différentiellement exprimée (DE), ou non significativement différentiée (n.s). Les croix représentent les abondances (métagénomique) des 1 404 fonctions non quantifiées en métatranscriptomique. Les triangles représentent les expressions (métatranscriptomique) des 9 fonctions non quantifiées en métagénomique.
metaG : métagénomique, metaT : métatranscriptomique.

Pour le second cas, nous avons cherché à vérifier que la détection des fonctions différentiellement exprimées (et non différentiellement abondantes) n'étaient pas due à des biais techniques dus à la variabilité du nombre de lectures prises en compte par échantillon. Cette variabilité de séquences pourrait impliquer une faiblesse de l'expression, une variance trop importante, ou une prévalence trop faible. Ainsi, pour chaque omique et chaque fonction, quatre comparaisons (entre lignées au sein d'un régime ou entre régimes au sein d'une lignée) ont été réalisées permettant d'identifier le statut DA et DE, seulement DA, seulement DE ou non différencié de chaque fonction dans chaque comparaison.

* ADN reliques : molécules d'ADN extra-cellulaires provenant de la dégradation de cellules mortes

En ce qui concerne l'impact de l'abondance/expression dans la sensibilité des analyses différentielles, nous avons évalué l'abondance/expression la plus forte parmi les deux groupes « lignée x régime » comparés. L'idée est de vérifier que les fonctions seulement DE ne sont pas en particulier dues à des comparaisons entre très faible expression. La **Figure F-5** montre que les expressions les plus fortes sont globalement moins importantes que les abondances les plus fortes. Pour autant, les fonctions qui ne sont détectées que différentiellement exprimées représentent globalement des fonctions pour lesquelles les quantifications sont globalement les plus importantes.

En ce qui concerne les écart-types des quantifications au sein de chaque groupe « lignée x régime », ils sont globalement très importants et plus importants pour les métatranscriptomes (±70,1%) que pour les métagénomes (± 53,9%), mais les fonctions qui sont détectées uniquement différentiellement exprimées représentent globalement les fonctions dont les quantifications sont les plus homogènes (±50,2% en métatranscriptomique et ±29,4% en métagénomique).

Enfin, nous avons contrôlé la prévalence. Celle des métatranscriptomes est légèrement moins importante (quantification moyenne dans 88% des échantillons) que celle des métagénomes (en moyenne dans 92,1%), mais encore une fois, ce sont les fonctions uniquement différentiellement exprimées qui présentent les prévalences les plus importantes.

Avec un nombre de fonctions différentiellement exprimées plus important que le nombre de fonctions différentiellement abondantes, les métatranscriptomes sont plus variables que les métagénomes. Ils ont également des mesures d'expression relative et de prévalence plus faibles, et de plus forts écart-types d'expression au sein des groupes « lignée x régime ». Pour autant, ces trois critères ne permettent pas de justifier cette forte détection de fonctions uniquement différentiellement exprimées. En effet, comparé au profil d'expression globale, celles-ci sont globalement bien exprimées avec une plus faible variabilité individuelle et avec une meilleure prévalence. Ces fonctions reflètent donc vraisemblablement une réalité biologique d'une plus grande sensibilité de l'activité transcriptionnelle du microbiote, en particulier suite à la modification du régime.

Ces analyses différentielles suggèrent donc que l'analyse du métagénome ne suffit pas pour interpréter fonctionnellement l'activité du microbiote et ses évolutions en fonction de son environnement. Au contraire, il pourrait porter à de fausses interprétations au vu du nombre non négligeable de fonctions abondantes mais non exprimées (1 404 fonctions) et de fonctions différentiellement abondantes mais non différentiellement exprimées (1 618 comparaisons).

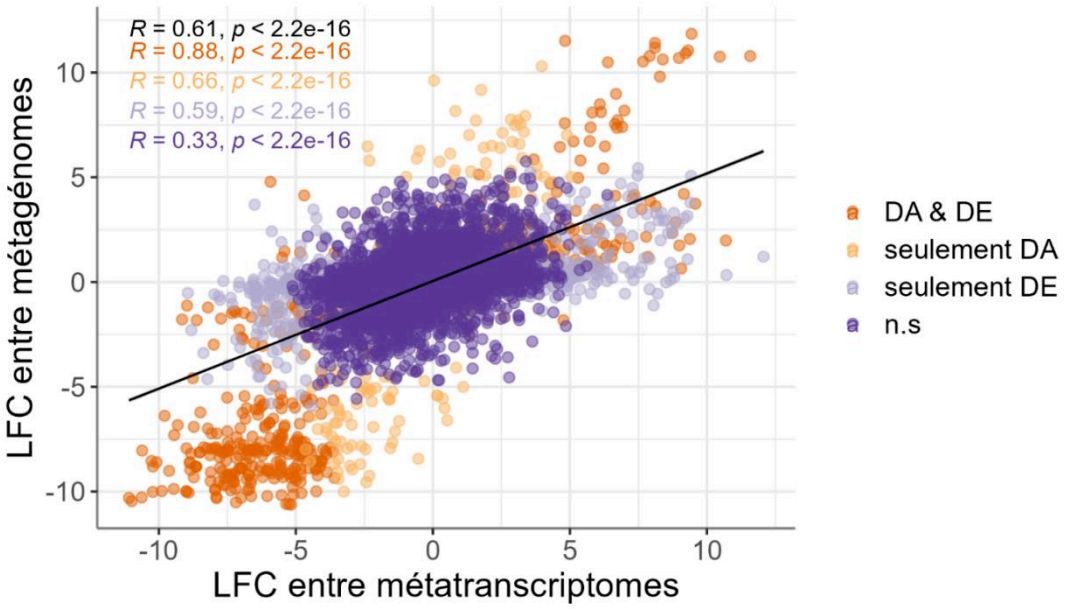


Figure F-6: Distribution des LFC entre métagénomes et métatranscriptomes.
LFC : « log fold change » niveau du rapport d'abondance ou d'expression des fonctions détectées différentiellement abondantes (DA) entre métagénomes et/ou différentiellement exprimées (DE) entre métatranscriptomes, ainsi que des fonctions non significativement variantes (n.s).

Pour autant, l'amplitude et le sens des différences d'abondances (« log fold change », LFC) des métagénomes selon la lignée ou le régime sont globalement positivement corrélés (R= 0,61) à ceux des différences d'expression des métatranscriptomes pour les mêmes comparaisons (**Figure F-6**). Ils peuvent même être très fortement corrélées si on regarde uniquement les fonctions qui sont différentiellement abondantes et exprimées (R= 0,88).

F.3.2. Analyse intégrative des profils d'abondances et d'expression

Pour mieux quantifier la similarité entre ces deux groupes de données omiques, nous avons réalisé une analyse intégrative (grâce des analyses triadiques partielles, ATP (Thioulouse and Chessel 1987; Girardie et al. 2024), en utilisant le package R ade4 (Dray, Dufour, and Chessel 2007)) des métagénomes et des métatranscriptomes en tenant compte des 34 échantillons et des 4 685 fonctions quantifiées dans les deux omiques (soit 1 404 fonctions quantifiées en métagénomique de moins et 9 fonctions quantifiées en métatranscriptomique de moins).

L'ATP permet une analyse commune de plusieurs tables dont les lignes et les colonnes correspondent aux mêmes variables (ici, les échantillons et les fonctions KEGG). Elle consiste notamment à calculer un coefficient RV (1) qui mesure l'interstructure, c'est-à-dire la ressemblance entre chaque paire de tables.

$$RV(X,Y) = \frac{\sum_{i=1}^{F} Cov(Xi,Yi)}{\sum_{i=1}^{F} \sqrt{Var(Xi) * Var(Yi)}}$$

(1) : équation du calcul du RV entre deux tables X et Y contenant F colonnes.
Cov : covariance ; Var : variance.

Ce coefficient RV est compris entre -1 et 1 et s'interprète comme un coefficient de corrélation. L'ATP calcule également un compromis pondéré entre l'ensemble des tables qui permet d'observer le comportement moyen entre les variables.

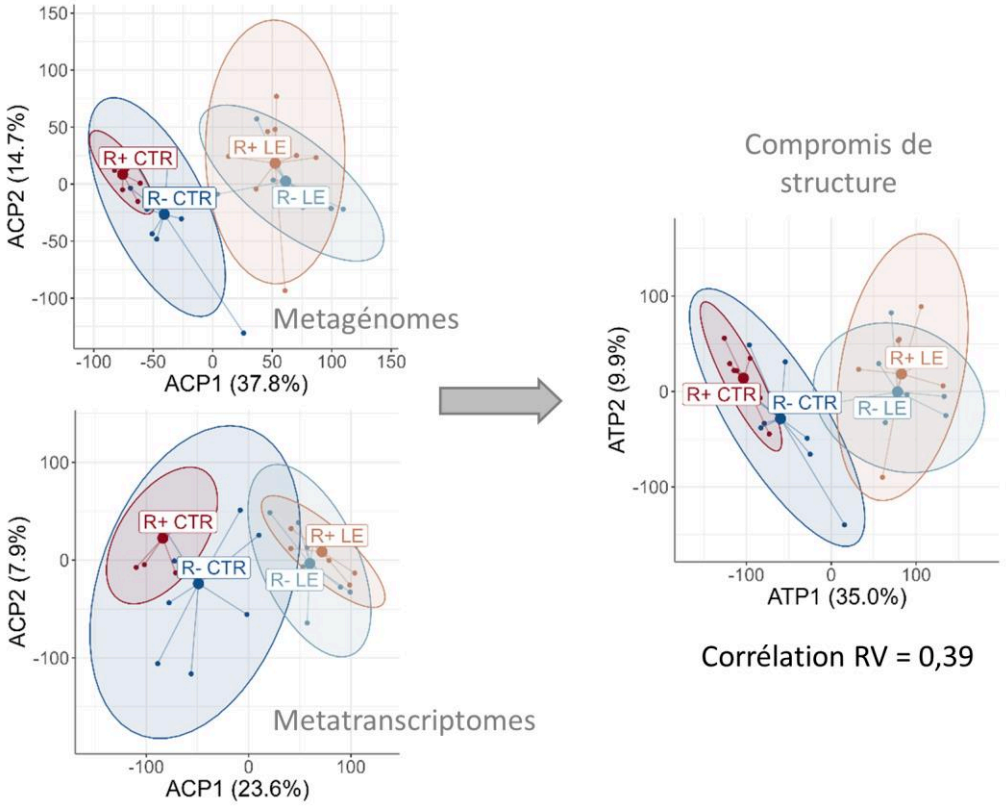


Figure F-7 : Analyse triadique partielle sans a priori sur la structuration des échantillons en fonction de la lignée et du régime.
Les analyses en composantes principales (ACP) sont réalisées sur les quantifications relatives normalisées CLR.

La construction du compromis peut se faire sans a priori sur l'appartenance des échantillons à des groupes distincts (ici « lignée x régime »), c'est-à-dire à partir d'analyses en composantes principales

(ACP, **Figure F-7**) ; ou bien en tenant compte d'une structuration des échantillons (ici en fonction de la lignée et du régime), c'est-à-dire en réalisant une analyse entre classe (« Between-Class Analysis », BCA, **Figure annexe J.2-1**) ; et enfin en ignorant la structuration des échantillons en fonction de la lignée et du régime, c'est-à-dire en réalisant une analyse intra-classe (« Within-Class Analysis », WCA, **Figure annexe J.2-2**).

Le **Tableau F-3** indique que les coefficients de corrélation RV sont modérés à forts selon que l'on prenne en compte ou non une structuration des échantillons en fonction de la lignée et du régime. Par ailleurs, il reste positif bien que plus faible (0,27), lorsque l'on élimine la structuration des échantillons en fonction de leur appartenance à une lignée et un régime.

Tableau F-3 : Coefficient RV entre métagénomes et métatranscriptomes.

	Coefficient RV
Sans apriori (ACP)	0,39
Entre groupe « lignée x régime » (BCA)	0,63
Intra groupe « lignée x régime » (WCA)	0,27

Les coefficients RV sont tous significatifs avec une *P*-value < 0,001. La significativité d'un RV est calculée par comparaison de la valeur observée à 9 999 autres valeurs de RV calculées après permutation des fonction d'une table (métatranscriptomes). La *P*-value correspond à la proportion des valeurs de RV permutés supérieures à la valeur de RV observée.

Il existe donc une association forte entre les métagénomes et les métatranscriptomes, portée par une structuration commune des échantillons en fonction de la lignée et du régime. A l'échelles des fonctions KEGG, les métatranscriptomes et les métagénomes présentent des corrélations majoritairement positives (moyenne à 0,3) s'échelonnant entre -0,47 et 0,99. Par ailleurs, les fonctions détectées différentiellement abondantes et différentiellement exprimées (DA & DE) ont les corrélations les plus fortes (moyennes à 0,56) (**Figure F-8**), corrélations qui sont, en proportion, les plus significatives. Les fonctions DA et DE contribuent donc le plus à la ressemblance entre omiques en contribuant probablement le plus à la structuration des échantillons selon la lignée et le régime.

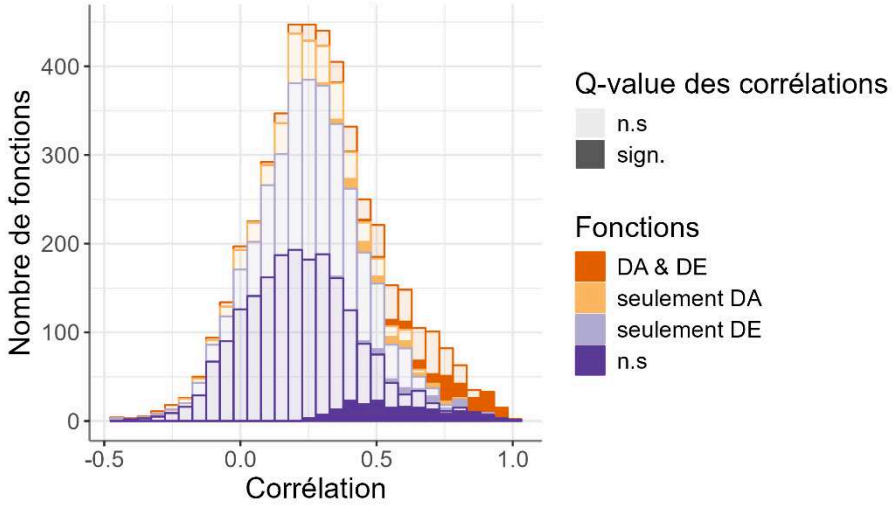


Figure F-8: Distribution des corrélations entre métagénomes et métatranscriptomes.
La significativité des corrélations est testée par 9 999 permutations des fonctions des métatranscriptomes et est corrigée par la méthode de Benjamini-Hochberg (Benjamini and Hochberg 1995). Le seuil de significativité est fixé à 0,1 qui équivaut à 0,05 de part et d'autre d'une distribution symétrique (George and Mudholkar 1990).

Alors que les analyses différentielles semblaient nous montrer de forte différences entre les métagénomes et les métatranscriptomes, la distribution des LFC et l'analyse intégrative, ATP, mettent en évidence une corrélation importante des mesures d'abondance et d'expression des fonctions. Les abondances des fonctions contribuent donc à structurer les expressions des fonctions du microbiote en fonction de la lignée et du régime.

F.4. Exploration des fonctions du microbiote cæcal différentiellement exprimées entre lignées et/ou régimes alimentaires

F.4.1. Mise en évidence des voies métaboliques et catégories fonctionnelles les plus différenciées

Pour explorer les nombreuses fonctions dont l'expression est influencée par la lignée ou le régime, nous nous sommes focalisés sur les voies métaboliques impliquant les fonctions qui contribuaient le plus à la structuration des métatranscriptomes en fonction de ces deux facteurs, grâce à des analyses de redondance (RDA (Rao 1964; Legendre and Legendre 2012), package R vegan (Oksanen et al. 2022)).

La RDA est une méthode multivariée qui combine l'ACP et les approches liées au modèle linéaire (régression, analyse de variance). Elle modélise l'effet d'une matrice de variables explicatives X sur une

matrice de variables réponse Y. Comme dans une ACP, la RDA calcule des axes orthogonaux qui « expliquent » le mieux la variation de Y ; mais elle « contraint » ces axes à être des combinaisons linéaires des variables de la matrice explicative X. Les inerties des axes de la RDA mesure l'inertie « expliquée » par le modèle. La RDA partielle est une extension de la RDA qui permet de tenir compte des covariables. Dans la RDA partielle, on étudie les effets de la matrice X sur la matrice Y, en les conditionnant, pour les effets d'une matrice de covariables W. Cela permet de quantifier l'effet d'un ensemble de variables explicatives, en les conditionnant pour un ensemble de covariables.

Dans notre étude, l'analyse de redondance globale repose sur un modèle linéaire incluant la lignée et le régime comme variables explicatives (l'interaction n'étant pas significative) appliqué à l'ensemble des 40 échantillons et des 4 694 fonctions exprimées. En conditionnant, la variable « lignée », nous identifions les fonctions qui contribuent le plus à la variable régime et inversement, en conditionnant la variable « régime », nous identifions les fonctions qui contribuent le plus à la variable « lignée ». De plus, à l'image des analyses différentielles, nous avons également procéder à des analyses reposant sur un modèle incluant uniquement la lignée ou uniquement le régime appliqué respectivement aux échantillons d'un seul régime ou d'une seule lignée (**Tableau F-4**). Ceci permet de trouver des fonctions qui contribuent à l'effet de la ligne ou du régime spécifiquement dans un régime ou une lignée.

Tableau F-4 : RDA : variance expliquée par les modèles contraints et significativité des variables explicatives, lignée et/ou régime.

	Global	**Comparaisons des lignées**		**Comparaisons des régimes**	
		Au sein du régime LE	Au sein du régime CTR	Au sein de la lignée R+	Au sein de la lignée R-
Variance expliquée par le modèle (%)	18,87	0,59	4,39	24,62	13,51
Anova					
Line	0,039*	0,234	< 0,001*		
Diet	< 0,001*			< 0,001*	< 0,001*

Le modèle global inclut `la lignée + le régime`. Les autres modèles incluent uniquement `la lignée` ou uniquement `le régime` et sont appliqués respectivement aux échantillons d'un seul régime ou d'une seule lignée. Une anova est utilisée pour mesurer l'effet de la lignée et/ou du régime sur les modèles contraints ; les astérisques mettent en évidence les effets significatifs (*P*-value < 0,05).

Ces analyses de redondances sont cohérentes avec les précédents résultats. Elles indiquent un effet fort du régime sur les métatranscriptomes des deux lignées, avec un effet plus marqué chez la lignée non efficiente R+ (qui explique 24,62% de la variance contre 13,51% chez la lignée efficiente R-), et un effet plus faible de la lignée qui n'est significatif que lorsque les poules sont nourries avec le régime

CTR (4,39% de variance expliquée par la lignée). Le modèle analysant l'effet de la lignée sous le régime LE n'étant pas significatif, il est exclu de la suite de l'analyse.

Pour chacun des autres modèles, la contribution à la structuration des échantillons selon la lignée et/ou le régime des 4 694 fonctions exprimées étant très progressive, nous avons choisi de sélectionner arbitrairement les 5% des fonctions (soit 235 fonctions) qui contribuaient le plus aux effets de chacun des facteurs, lignée ou régime. Par rapport à une contribution moyenne de 0,0213%, ces 235 fonctions contribuent de 5,2 à 49 fois plus. Parmi ces fonctions, nous avons vérifié le nombre de fonctions différentiellement exprimées (DE) et/ou différentiellement abondantes (DA, sur les métagénomes) (**Tableau F-5**).

Tableau F-5 : Nombre de fonctions DA et/ou DE parmi le top 5% des fonctions qui contribuent le plus à la structuration des métatranscriptomes en fonction de la lignée et du régime.

	Comparaisons des lignées		**Comparaisons des régimes**		
Modèle	Global	Au sein du régime CTR	Global	Au sein de la lignée R+	Au sein de la lignée R-
Nombre de fonction DE	79	94	233	234	226
Nombre de fonction DA	29	42	165	139	130
Nombre de fonction DA & DE	15 (18,9%)	21 (22,3%)	164 (70,4%)	139 (59,4%)	128 (56,6%)

Le top 5% des fonctions par modèle correspond à 235 fonctions. Le modèle global correspond à `lignée + Condition(le régime)` ou `régime + Condition(lignée)`. Le pourcentage des fonctions DA & DE est exprimé en fonction du nombre de fonctions DE. DE : différentiellement exprimée ; DA : différentiellement abondante.

Ainsi, parmi les 235 fonctions qui contribuent à la séparation des métatranscriptomes selon le régime alimentaire, la quasi-totalité sont différentiellement exprimées et, selon le modèle, de 56,6% à 70,4% sont également différentiellement abondantes. Ceci est en cohérence avec le fait que le régime alimentaire a un impact modéré à fort selon la lignée sur les métatranscriptomes (*P*-value < 0,001 et variance expliquée entre 13,5 et 24,6%). Cela montre également que l'abondance des gènes et donc des fonctions joue un rôle important dans la mesure de leurs expressions et dans leurs sensibilités au régime. Au contraire, parmi les 235 fonctions qui contribuent à la séparation des métatranscriptomes selon la lignée, une minorité de fonctions (79 pour le modèle global et 94 pour le modèle au sein du régime CTR) sont différentiellement exprimées et environ 20% d'entre elles sont également différentiellement abondantes. La lignée a donc un effet faible (*P*-value globale de 0,039 proche du seuil de significativité et variance expliquée maximum de 4,4%) et elle influence l'expression des fonctions indépendamment de leur abondance.

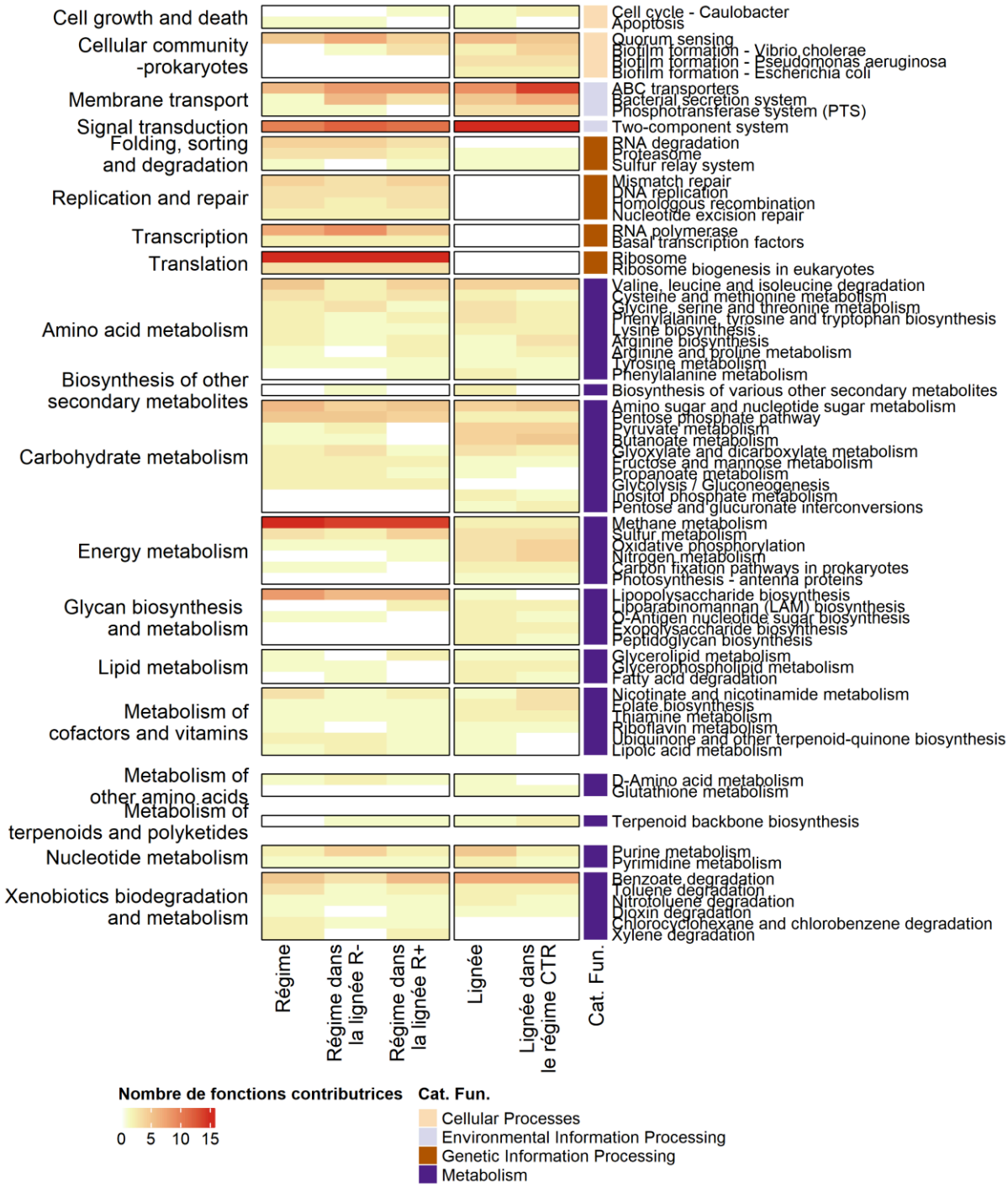


Figure F-9 : Catégories fonctionnelles structurant les métatranscriptomes.
Heatmap du nombre de fonction de 72 catégories fonctionnelles (dont 51 voies métaboliques) avec au moins 2 fonctions parmi les 5% des fonctions qui contribuent le plus à la structuration des métatranscriptomes en fonction de la lignée ou du régime. La classe « Human diseases » est exclue. Les colonnes « Régime » et « Lignée » correspondent au modèle linéaire global incluant la lignée et le régime, les autres colonnes correspondent aux modèles incluant la lignée ou le régime appliqués respectivement aux échantillons d'un régime ou d'une lignée.

Au total, 575 fonctions permettent d'identifier 185 catégories fonctionnelles KEGG (dont 87 voies métaboliques) potentiellement influencées par la lignée et/ou le régime.

La **Figure F-9** met en particulier en évidence que les métatranscriptomes des deux lignées et des deux régimes se différencient sur un ensemble de fonctions qui permettent aux micro-organismes d'interagir avec leur environnement (détection du quorum, transduction de signaux, transport transmembranaire). Par ailleurs, la catégorie fonctionnelle liée aux activités de transcription et traduction de l'information génétique montre également des signaux forts de différenciation des métatranscriptomes entre les deux régimes. Enfin, on peut noter que de nombreuses fonctions contribuant à la différenciation des métatranscriptomes selon la lignée et/ou le régime interviennent dans des métabolismes variés des carbohydrates (les sucres aminés (osamines) et nucléotidiques (nucléosides), pentose phosphate, pyruvate, ou butanoate), de l'énergie (méthane, ou nitrogène) ou encore dans la formation de lipopolysaccharides ou la dégradation du benzoate.

F.4.2. Influence de caractéristiques physiques et génétiques de différents types de micro-organismes

Parmi les métabolismes et autres catégories fonctionnelles influencées par la lignée ou le régime alimentaire, nombreux sont ceux qui reflètent en réalité une différenciation de composition microbienne. En effet, bien que de nombreux processus cellulaires soient retrouvés dans tous les royaumes taxonomiques, les mécanismes et donc les fonctions impliquées peuvent différer. C'est le cas notamment des processus de bases liés à l'ADN, l'ARN ou aux protéines (catégorie fonctionnelle « *Genetic Information Processing* »). Dans notre étude, ces processus sont composés de fonctions dont l'expression contribue à la différenciation des métatranscriptomes selon le régime. En l'occurrence, la quasi-totalité des fonctions impliquées dans ces processus et qui sont surexprimées dans la condition du régime LE est codée par des gènes affiliés aux archées et dans une moindre mesure à des gènes eucaryotes. Alors que les fonctions pour ces mêmes processus, qui sont surexprimées dans la condition du régime CTR, regroupent des gènes affiliés à 98,5% à des bactéries. Si l'on se reporte aux analyses différentielles réalisées à l'échelles des MGS, nous avons effectivement détectée une MGS affiliée aux archées (genre *Methanomassiliicoccus_A* du phyla Thermoplasmatota). Cette MGS est absente dans les microbiotes des poules nourries avec le régime CTR et a une abondance relativement élevée de 0,1% dans les microbiotes des poules nourries avec le régime LE, et ce quel que soit la lignée de poule. A l'échelle des gènes, on observe une diversité taxonomique des archées plus importante avec notamment deux autres genres (*Methanobrevibacter* et *Methanocorpusculum*) appartenant à un autre phylum (Euryarchaeota). Le schéma de présence/absence des archées selon le régime semble contribuer de façon importante à la structuration des métatranscriptomes et provoque la mise en évidence des fonctions génétiques de bases liées au type de micro-organisme. Notre étude se

focalisant sur les procaryotes, nous ne pouvons explorer plus en détail la différenciation de ces processus selon les différences d'abondance des eucaryotes.

L'influence de la différence d'abondance des archées ne s'illustre pas seulement à l'échelle de ces processus de traitement de l'information génétiques. En effet, le métabolisme du méthane fait partie des métabolismes dont l'expression des fonctions est largement impactée par le régime alimentaire. La production de méthane par le microbiote intestinal est majoritairement étudiée chez les ruminants car ils participent de façon significative aux émissions de gaz à effet de serre (Tapio et al. 2017). La production de méthane intervient lors de la fermentation, qui augmente avec la quantité d'aliments non digestes par l'hôte, tels que la cellulose ou les hémicelluloses, comme cela est le cas dans notre régime LE. Cette fermentation produit, en plus des métabolites comme les SCFA (acides gras à chaîne courte), du dihydrogène qui, lorsqu'il s'accumule, réduit l'activité de fermentation (Cisek et al. 2023). Parmi les micro-organismes capables de consommer ce dihydrogène se trouve les archées dites méthanogènes, c'est-à-dire qui produisent du méthane. Même si les méthanogènes contribuent au maintien d'un environnement propice à la fermentation et donc par extension à la production de métabolites bénéfiques pour l'hôte, elles sont généralement considérées comme défavorables pour celui-ci. En effet, elles sont en compétition pour la consommation du dihydrogène avec d'autres bactéries, notamment les bactéries acétogènes, qui produisent l'acétate, un SCFA source d'énergie pour une multitude de types cellulaires de l'hôte (Misiukiewicz et al. 2021; Cisek et al. 2023). Il existe trois catégories de méthanogènes qui se différencient selon le substrat utilisé en combinaison avec le dihydrogène : les hydrogénotrophes utilisent le CO2 et le formate ; les méthylotrophes utilisent des composés méthylés comme le méthanol, ou les méthylamines ; et les acétotrophes (ou acétoclastiques) utilisent l'acétate (Misiukiewicz et al. 2021; Aryee et al. 2023).

La **Figure F-10** illustre que dans notre étude, parmi ces trois voies possibles de production de méthane, la voie méthylotrophe est en surexpression dans les microbiotes de la condition du régime LE quelle que soit la lignée, et cette voie correspond en effet à celle utilisée par *Methanomassiliicoccus_A* (la MGS identifiée en surabondance dans la condition de régime LE). Par ailleurs, la voie hydrogénotrophe est également partiellement représentée par 2 fonctions supplémentaires en surexpression dans la condition du régime LE chez la lignée R+, suggérant la présence d'autres archées. En regardant les taxonomies associées aux gènes codant les fonctions de ces 2 voies et en surexpression dans la condition de régime LE, on identifie une autre famille d'archées, *Methanomethylophilaceae,* appartenant au même phylum Thermoplasmatota et également méthylotrophe (Gaci et al. 2014), et l'espèce *Methanobrevibacter woesei*, du phylum Euryarchaeota qui est hydrogénotrophe (Misiukiewicz et al. 2021). Les *Methanobrevibacter* sont généralement les archées les plus abondantes

du microbiote intestinal chez une multitude d'espèce hôtes ruminantes ou non (Moissl-Eichinger et al. 2018; Misiukiewicz et al. 2021) dont la poule (Saengkerdsub et al. 2007; Aruwa et al. 2021; J. Yang et al. 2022), mais la diversité des méthanogènes est également fortement impactée par l'environnement et notamment le régime (Luo et al. 2017; Cisek et al. 2023). Ainsi Cisek et al. ont identifié pour la première fois dans le microbiote de poule le genre *Methanomassiliicoccus*, et à l'image de notre étude, il était le genre dominant des méthanogènes.

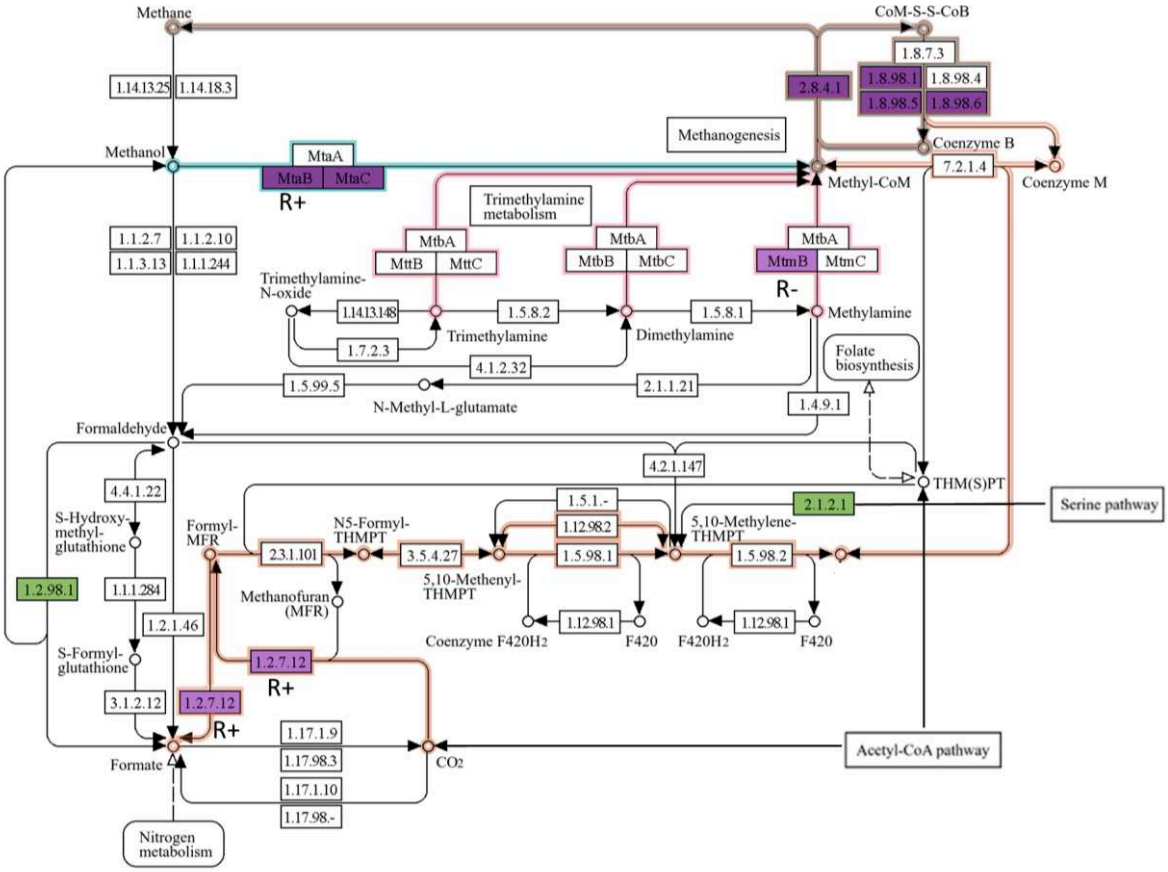


Figure F-10 : Extrait du métabolisme du méthane et fonctions différentiellement exprimées entre régimes.
Les voies méthylotrophiques indiquées par les flèches, sont mises en surbrillance bleu et rose, et la voie hydrogénotrophique est en surbrillance orange. Les fonctions surexprimées dans la condition de régime LE sont représentées par les rectangles violet (uniquement dans une lignée si précisée), et celles surexprimées dans la condition de régime CTR par les rectangles vert. Les rectangles violet foncé représentent les fonctions qui sont également parmi le top 5% des fonctions contributrices à la différenciation des métatranscriptomes entre régime. La carte du métabolisme du méthane est adaptée de celle produite par KEGG (Kanehisa et al. 2016).

Les différences d'abondance des archées ne sont pas les seules différences liées à la composition microbienne qui mettent particulièrement en évidence certaines catégories fonctionnelles ou métaboliques. Par exemple, le métabolisme de la biosynthèse des lipopolysaccharides (LPS),

constituants majeurs de la membrane externe des bactéries Gram négatives (c'est-à-dire avec une double membrane), est majoritairement représenté par des fonctions surexprimées dans la condition du régime CTR dans les deux lignées, en particulier dans la lignée R- (avec 16 fonctions) et de façon plus modérée dans la lignée R+ (7 fonctions). Ces LPS sont produits à partir de composés issus du métabolisme du pentose phosphate, des sucres aminés et nucléotidiques, et des sucres antigènes O (domaine immunogène des LPS). Les fonctions de ces trois voies métaboliques sont d'ailleurs très majoritairement en surexpression dans les microbiotes de la condition de régime CTR dans les deux lignées. Mêmes si ces voies sont centrales et à l'interface avec de multiples autres voies métaboliques que celle des LPS, une partie des fonctions en surexpression dans la condition de régime CTR pourraient également être expliquées par une potentielle surabondance de bactéries Gram négatives dans cette condition. Toutefois, nous manquons d'annotation sur le type de Gram de chaque bactérie pour confirmer cette hypothèse.

Enfin le métabolisme du butyrate (ou butanoate) est mis en évidence par les analyses de redondance comme contribuant à la différenciation des métatranscriptomes en fonction de la lignée des poules. Ce métabolisme est représenté par des fonctions surexprimées presque exclusivement chez la lignée non efficiente R+ (5 fonctions sur 6 DE). Sur ces 5 fonctions, 3 permettent de couvrir la quasi-totalité des fonctions permettant de produire du butane-2,3-diol et de l'acétoïne, produits de la fermentation du glucose. L'acétoïne a longtemps été utilisée en microbiologie, pour identifier la présence d'entérobactéries (Dart 1996; Campbell, Waud, and Matthews 2009). Parmi les MGS quantifiées dans notre étude, une appartient à la famille des Enterobacteriaceae (espèce *Escherichia coli*) et est effectivement en plus grande abondance dans les microbiotes de la lignée R+ que dans ceux de la lignée R- lorsque les poules sont nourries avec le régime CTR. Alors que le métabolisme du butyrate supposait une différenciation dans la production de cet acide gras particulièrement important pour les cellules épithéliales de l'intestin, les analyses des métatranscriptomes ne permettent pas de mettre en évidence une réelle surexpression des fonctions aboutissant directement à une surproduction de cet SCFA chez la lignée efficiente R+. Pourtant cette surproduction de butyrate tend à être observée sur notre expérience de dosage des acides gras (Tableau 3 et figure supplémentaire S1 de l'article publié, paragraphe D.5).

F.4.3. Le transport et le métabolisme des carbohydrates tendent à illustrer les différences d'activités du microbiote entre lignées et entre régimes

Les analyses de métatranscriptomique mettent également en évidence de nouvelles pistes, en particulier grâce aux différences d'expression des fonctions liées aux transporteurs ABC (« ATP-Binding

Cassette »). Ces transporteurs utilisent l'ATP comme énergie pour permettre le passage au travers de la membrane cellulaire de multiples substrats comme les sucres, les ions, les minéraux, les vitamines, les lipides ou encore les acides aminés (Thomas and Tampé 2020; Kanehisa Laboratories, n.d.). L'expression des fonctions liées à ces transporteurs permet à la fois de différencier les microbiotes des deux lignées (en condition de régime CTR, **Figure F-11**) et les microbiotes en fonction des deux régimes alimentaires (quelle que soit la lignée).

La différenciation des microbiotes en fonction de la lignée se joue notamment au niveau du transport des sucres. Les microbiotes de la lignée R+ surexpriment des fonctions permettant le transport de sucres complexes comme l'arabinogalactane et le malto-oligosaccharide, ou le raffinose, le stachyose et le melibiose. *A contrario*, les microbiotes de la lignée R- surexpriment des fonctions permettant le transport de sucres simples comme L-arabinose, D-allose ou l'inositol. Un autre système de transport des sucres est mis en évidences par les analyses de redondance, le système phosphotransférase (PTS), qui utilise le phosphoenolpyruvate comme énergie. Deux transporteurs sont codés par des fonctions surexprimées dans les microbiotes de la lignée non efficiente R+, ceux du maltose (constitué de glucose) et du galactitol (dérivé du galactose). Par ailleurs, un transporteur est composé de fonctions surexprimées dans le microbiote de la lignée efficiente R-, permettant à la fois le passage du maltose et du glucose.

La différenciation des métatranscriptomes en fonction du régime alimentaire reprend la majorité des différences d'expression des fonctions liées à ces précédents transporteurs. Les précédentes fonctions surexprimées dans la lignée R+ sont également surexprimées dans la condition de régime CTR, et une partie des fonctions surexprimées dans la lignée R- sont également surexprimées dans la condition de régime LE. De plus, de nombreux autres transporteurs de sucres mais également d'acides aminés et de minéraux sont également mis en évidence par des différences d'expressions des fonctions des microbiotes selon le régime.

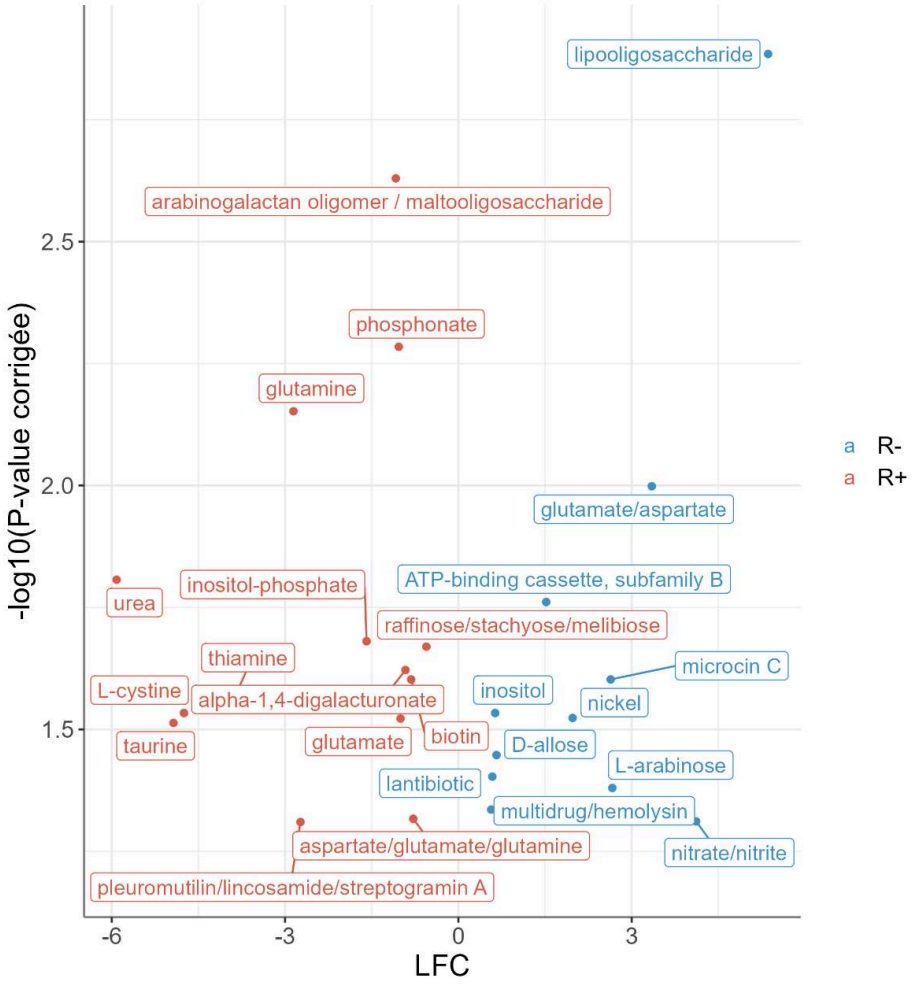


Figure F-11 : Transporteurs ABC dont les fonctions sont différentiellement exprimées entre lignées au sein du régime CTR.
Le modèle compare R- à R+, les « log fold change » (LFC) obtenus indiquent une surexpression dans R- s'ils sont positifs (bleu) et une surexpression dans R+ s'ils sont négatifs (rouge). A l'échelle des transporteurs ABC, le LFC et la *P*-value corrigée représentent les valeurs moyennes des LFC et *P*-value ajustées des fonctions différentiellement exprimées de chaque transporteur.

Ces résultats suggèrent notamment de potentielles différences dans le métabolisme de différents carbohydrates et donc une différence d'activité des CAZymes. En effet, les CAZymes sont les principales enzymes qui permettent notamment de dégrader les sucres complexes en sucres plus simples qui interviendront ensuite dans différents métabolismes, notamment celui de la synthèse de SCFA. Ces enzymes fonctionnent en association avec d'autres gènes liés physiquement sur le génome (typiquement en opéron) et forment des clusters de gènes de CAZymes (« CAZyme Gene Clusters », CGC) (H. Zhang et al. 2018). Les CGC se composent de trois catégories de gènes, au moins une CAZyme, un transporteur, et un facteur de transcription. Les transporteurs ABC et les PTS constituent des types de transporteurs couramment rencontrés dans ces CGC en particulier chez les bactéries Gram positives (Bedu-Ferrari et al. 2022; Z. Wang et al. 2022).

Notre précédente étude fonctionnelle (basée sur de l'inférence de fonctions à partir du séquençage métabarcoding, chapitre **D**) nous a permis d'émettre l'hypothèse d'une différenciation entre lignée (nourries avec le régime CTR) à l'échelle du métabolisme de l'amidon et du saccharose et du métabolisme du propionate (confirmé par l'expérience de dosage des SCFA). Bien que ces métabolismes n'aient pas été identifiés par les analyses de redondances comme structurant les métatranscriptomes en fonction de la lignée, nous avons voulu vérifier leur comportement à l'échelle des fonctions différentiellement exprimées.

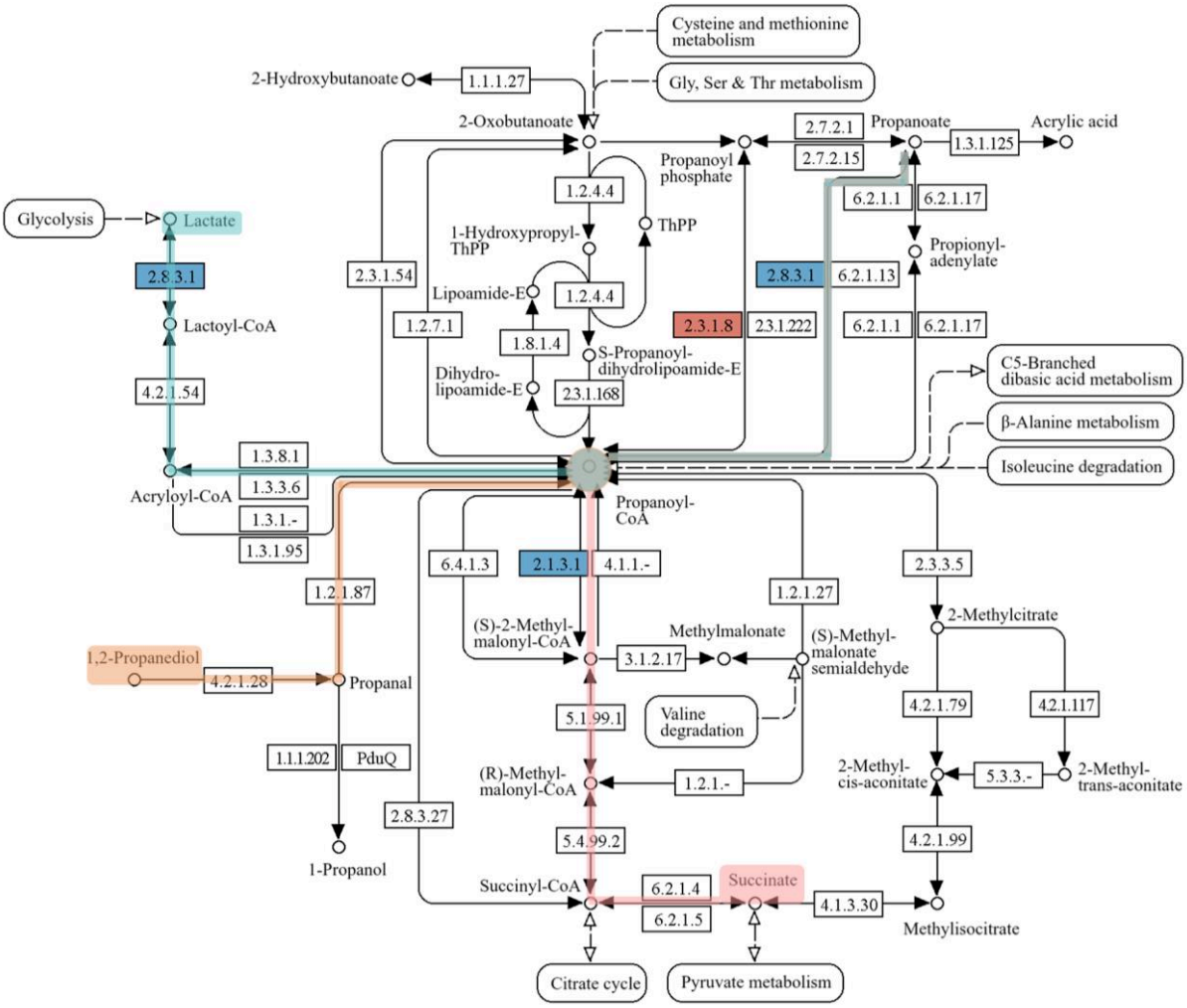


Figure F-12 : Fonctions du métabolisme du propionate différentiellement exprimées entre lignée.
Les expressions selon la lignée sont comparées lorsque les poules sont nourries avec le régime CTR. Les fonctions en bleu sont détectées en surexpression dans la lignée efficiente R-, les fonctions en rouge sont détectées en surexpression dans la lignée non efficiente R+. La carte du métabolisme est adaptée de celle produite par KEGG (Kanehisa et al. 2016).

Le propionate est produit selon trois voies différentes, à partir du succinate (généralement la voie majoritairement observée), du propanediol ou du lactate (Reichardt et al. 2014) (**Figure F-12**). Selon

l'analyse de métabarcoding, les fonctions de ces 3 voies étaient en surabondance dans le microbiote de la lignée efficiente R- (**Figure annexe J.2-3**), lignée dans laquelle le propionate est effectivement produit en plus forte proportion. A partir des métatranscriptomes, le nombre de fonction DE est fortement diminué (7 fonctions en surabondances chez la lignée R- sur les données de métabarcoding contre 2 en surexpression sur les données de métatranscriptomique). Ces deux fonctions se positionnent sur les voies du succinate et du lactate (**Figure F-12**), mais elles ne les couvrent que partiellement, ne permettant pas de conclure sur la réalité de la surexpression de ces voies métaboliques.

De même que pour le métabolisme du propionate, le nombre de fonctions DE impliquées dans le métabolisme de l'amidon et du saccharose est nettement moindre que le nombre de fonctions DA détectées sur l'analyse de métabarcoding (5 DE contre 13 DA) et aucune d'entre elles n'est détectée dans les deux analyses. Ces 5 fonctions sont principalement plus exprimées dans le microbiote de la lignée R+, à l'image de l'analyse de métabarcoding, et tendent donc à confirmer une plus forte activité vers la dégradation de l'amidon, du saccharose et du maltose que chez la lignée R-, mais les voies de dégradation/synthèse de ces sucres n'étant que partiellement couvertes, cela doit être confirmé.

Ces différences entre analyse de métabarcoding et de métatranscriptomique, peuvent s'expliquer par des raisons biologiques et techniques. En effet, bien qu'une part de l'expression soit due à son abondance initiale dans le microbiote, il est également possible qu'un gène ne soit pas transcrit. L'analyse d'abondance inférée pourrait donc tenir compte de fonctions non actives en réalité. Par ailleurs, notre analyse des données de métabarcoding repose sur l'inférence de fonctions qui ne tient pas compte de l'ensemble des ASV (70% des ASVs correspondant à 91% des abondances, paragraphe **D.3**). De plus, le gène de l'ARNr 16S est présent en un nombre variable de copies selon les génomes bactériens. Bien que ce biais de quantification soit pris en compte par la suite d'outil PICRUSt2 pour corriger l'abondance des fonctions, il est tout à fait probable que ces abondances ne reflètent que partiellement la réalité. Enfin pour les données de métatranscriptomique, l'influence de la profondeur de séquençage joue sur la richesse des gènes pris en compte dans la mesure d'expression des fonctions, tout comme le taux d'annotation fonctionnelle des gènes.

F.5. Bilan des résultats et réflexion sur le développement méthodologique

Cette première analyse du métatranscriptome cæcal de poule, une des premières de la littérature scientifique a été faite sans *a priori*. Elle a ainsi permis de décrire un large panel de catégories

fonctionnelles dont l'activité est influencée par des caractéristiques de l'hôte (la lignée) et d'environnement d'élevage (le régime). Elle a également mis en évidence un plus grande sensibilité des métatranscriptomes que des métagénomes aux différentes conditions. Par ailleurs, l'abondance des fonctions contribue fortement dans la quantification de l'expression des fonctions. Ainsi, nous avons pu mettre en évidence que de nombreuses différences d'expressions reflètent notamment des différences d'abondances et donc *in fine* de certains micro-organismes. Enfin, elle a permis d'identifier quelques métabolismes particulièrement intéressant qu'ils convient d'explorer plus en détails. Ainsi, pour tenter de répondre plus précisément aux questions biologiques posées au démarrage de cette thèse, des analyses complémentaires seront réalisées en tenant compte de ces premiers résultats. Ces analyses incluront une normalisation des expressions des fonctions par leur abondance et l'analyse ciblée des CAZYmes et des fonctions associées (**G.3**).

Lors de cette analyse exploratoire de l'activité réelle des microbiotes, nous avons également cherché à relier certaines fonctions KEGG discriminant nos groupes d'échantillons et les micro-organismes porteurs de ces fonctions. Et, bien que seulement 24,8% des gènes soient annotés taxonomiquement de façon complète jusqu'à la famille (16,2% jusqu'au genre et 10,8% jusqu'à l'espèce), ils nous ont permis de supposer la présence de micro-organismes pour lesquels nous n'avons pas de MGS suffisamment couvertes et/ou abondantes, en particulier parmi les archées et les eucaryotes. Ainsi pour exploiter au mieux les différentes sources d'annotation taxonomique (en particulier des gènes), nous pourrions envisager :

- ajouter aux MGS les génomes présents dans les bases de référence des espèces supplémentaires identifiées par les gènes (par exemple GTDB propose un génome représentatif par espèce procaryote).
- appliquer la stratégie de constitution de MSP directement à partir de l'ensemble des gènes (sans *a priori* sur leur appartenance à un MAG) (Plaza Oñate et al. 2019) pour compléter la richesse des espèces représentées par les MGS.
- analyser les profils d'abondance et de prévalence des MGS associées aux taxonomies détectées par les gènes pour paramétrer les filtres sur les abondances et prévalences. Ceci permettrait de conserver des MGS rares mais dont nous avons suffisamment de preuves de leur présence au vu de leur abondance et prévalence sur les gènes.

En ce qui concerne les micro-organismes eucaryotes, comme dans la majorité des études sur les microbiotes intestinaux, ils sont ignorés de notre analyse par l'utilisation de la base d'affiliation taxonomique GTDB dédiée aux organismes procaryotes. Dues à leur plus grande taille et à la

complexité de leur génome, ils sont plus difficiles à assembler en contigs et à regrouper en MAG de bonne qualité (Massana and López-Escardó 2022; Alexander et al. 2023). Bien qu'ils soient attendus en très faible richesse et abondance dans les microbiotes intestinaux notamment de la poule (Huang et al. 2018; Hooks and O'Malley 2020), il pourrait être intéressant pour une suite logicielle comme metagWGS d'être exhaustive dans l'identification des micro-organismes que composent un milieu.

G. Productions, Valorisations, Perspectives

En utilisant trois techniques omiques de caractérisation d'un écosystème microbien, cette thèse visait à répondre à la fois à des objectifs scientifiques et techniques. Comme indiqué en introduction, ces techniques omiques se différencient notamment sur leur niveau de facilité de mise en œuvre. En effet, nous avons vu dans les chapitres précédents qu'elles ont nécessité, selon la méthode, des mises au point expérimentales et analytiques. Grâce à l'analyse des données de métabarcoding, des réponses ont pu être apportées aux problématiques biologiques de cette thèse, à savoir la caractérisation du lien entre microbiote cæcal et efficience alimentaire, et de l'impact du régime alimentaire sur cette association. Pour la métagénomique et la métatranscriptomique, ces mises au point ont été plus conséquentes. Bien que leurs interprétations taxonomiques et fonctionnelles ne soient pas finalisées à ce jour, cette étude nous a permis d'acquérir des compétences sur des protocoles expérimentaux, sur la mise en place de contrôle qualité, et sur des outils et procédures d'analyse. Les nombreuses réflexions sur la façon d'analyser ces séquences ont par ailleurs fait ou feront l'objet de valorisations en termes de développement bioinformatique. Elles nous ont également permis de mettre clairement en évidence quelques points de blocage que nous prévoyons de résoudre pour apporter de nouvelles réponses aux objectifs de cette thèse, notamment technique, à savoir la comparaison de ces trois techniques omiques.

En plus de l'article scientifique publié sur la base des résultats obtenus sur les données de métabarcoding (Bernard et al. 2024 et paragraphe **D.5**), les paragraphes qui suivent décrivent les valorisations réalisées ou à venir en termes de développement d'outils informatiques, ainsi que les publications scientifiques que nous prévoyons de réaliser incluant les données de métagénomique et de métatranscriptomique.

G.1. De l'utilisation des outils bioinformatiques au développement de nouvelles fonctionnalités

L'analyse des séquences de métabarcoding et de métagénomique s'est principalement reposée sur l'utilisation des suites logicielles FROGS (Escudié et al. 2018; Bernard et al. 2021) dont je suis co-développeuse, et metagWGS (Fourquet et al. 2022). En devenant utilisatrice de ces outils, j'ai pu mettre à contribution mon expérience acquise durant cette thèse pour suggérer et co-développer des améliorations. Les développements réalisés, en cours ou à venir concernent l'amélioration des

performances, l'ajout de nouvelles fonctionnalités, ou l'ajout d'indicateurs permettant d'accompagner l'utilisateur à la prise de décision (**Figure G-1**).

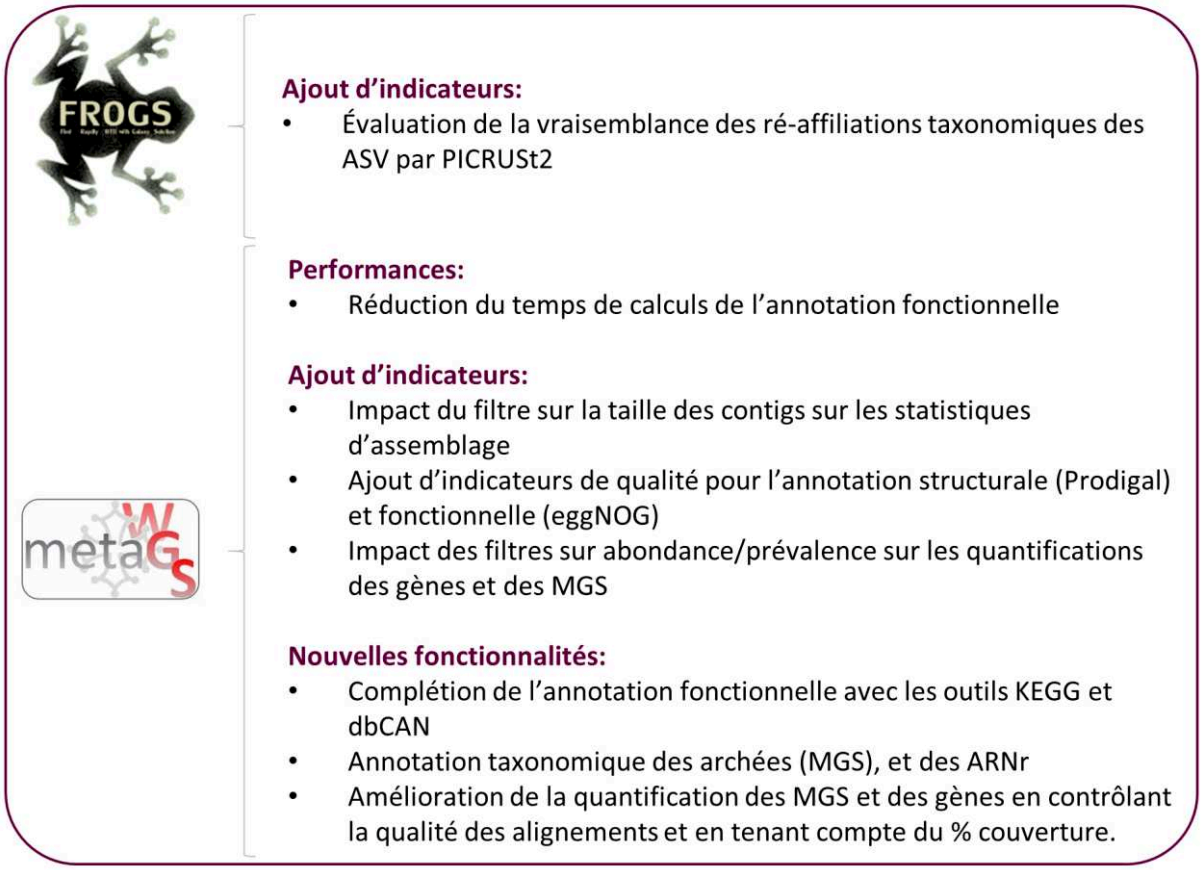


Figure G-1 : Contributions aux développements bioinformatiques des outils utilisés pendant la thèse. Ces développements ont été réalisés et mis à disposition de tous, sont en cours, ou en réflexion avec les développeurs de ces outils.

De plus, les nombreuses comparaisons de procédures d'analyses des données de métagénomiques, suggèrent des possibilités de nouveaux développements pour exploiter au mieux les séquences. En particulier, la discrimination des métagénomes en espèces distinctes ne semblent pas optimale, et pourrait être améliorée en optimisant la procédure de déréplication, ou en exploitant également l'information d'abondance des gènes pour construire des MSP (paragraphe **E.4**). Par ailleurs, les gènes permettent également d'apporter une information sur la diversité taxonomique des microbiotes. Celle-ci pourrait être exploitée pour compléter la richesse taxonomique détectée à l'échelle des MGS ou paramétrer les seuils de filtres sur l'abondance et la prévalence (paragraphe **F.5**). Ces suggestions de développement nécessitent toutefois plus de réflexions pour intégrer ces différentes sources d'annotations ou regroupement taxonomiques.

G.2. Finalisation et valorisation de l'analyse comparative des données omiques

Grâce au dispositif expérimental de cette thèse nous avons la possibilité de comparer trois méthodes de séquençage d'un écosystème microbien complexe, d'une part à l'échelle des taxonomies (comparaison du métabarcoding et de la métagénomique) et à l'échelle des fonctions (comparaison des trois techniques).

La comparaison des compositions microbiennes obtenues jusque-là à partir des données de métabarcoding et de métagénomiques nous ont montré des tendances similaires mais également des divergences : même influence de la lignée et du régime sur la diversité microbienne, même composition générale au phylum et cohérence partielle des genres différenciés entre condition. Elle confirme également la meilleure sensibilité de la métagénomique dans la détection des différents micro-organismes présents et une plus grande résolution taxonomique dans leur annotation. *A* contrario, la métagénomique est plus sensible aux variations de profondeur de séquençage qui est inévitablement plus élevée que celle nécessaire au métabarcoding et nécessite des mises au point analytiques plus importantes. Cependant, la comparaison exhaustive de ces deux techniques est rendue difficile par l'utilisation de deux référentiels taxonomiques différents. Pour résoudre cela, nous prévoyons d'affilier à nouveau les ASV et les MGS selon la classification NCBI (grâce aux séquences d'ARNr 16S de la base RefSeq pour les ASV, et aux taxonomies synonymes NCBI de la base GTDB pour les MGS), afin de poursuivre les comparaisons entre analyse des données de métabarcoding et des données de métagénomique *(de-novo* et enrichie avec le catalogue MetaChick), **Figure G-2-A**).

A l'échelle des fonctions du microbiote, la comparaison des profils métagénomiques et métatranscriptomiques a confirmé une plus grande stabilité des profils fonctionnels vis-à-vis des profils taxonomiques mais une plus grande variabilité des expressions comparées aux abondances des fonctions (diversité alpha et bêta). Les analyses différentielles d'abondance et d'expression (DA/DE) des fonctions, ainsi que l'analyse triadique partielle (ATP), nous ont toutefois permis de mettre en évidence une corrélation importante des profils d'abondances et d'expressions aboutissant notamment à une structuration commune des échantillons selon la lignée et/ou le régime. Parallèlement, nous avons cherché à comparer de façon non exhaustives (quelques métabolismes ciblés), les profils d'abondance obtenus sur les données de métabarcoding avec les profils d'expression obtenus sur les données de métatranscriptomique. Ces comparaisons n'ont montré que peu de chevauchements entre catégories fonctionnelles particulièrement différenciées entre lignées ou entre régimes. Pour avoir une comparaison plus exhaustive des profils fonctionnels et comparer également les profils d'abondance des données de métabarcoding et des données de métagénomique, nous

prévoyons de compléter l'analyse triadique partielle en incluant la table des profils d'abondance basés sur les données de métabarcoding (**Figure G-2-B**).

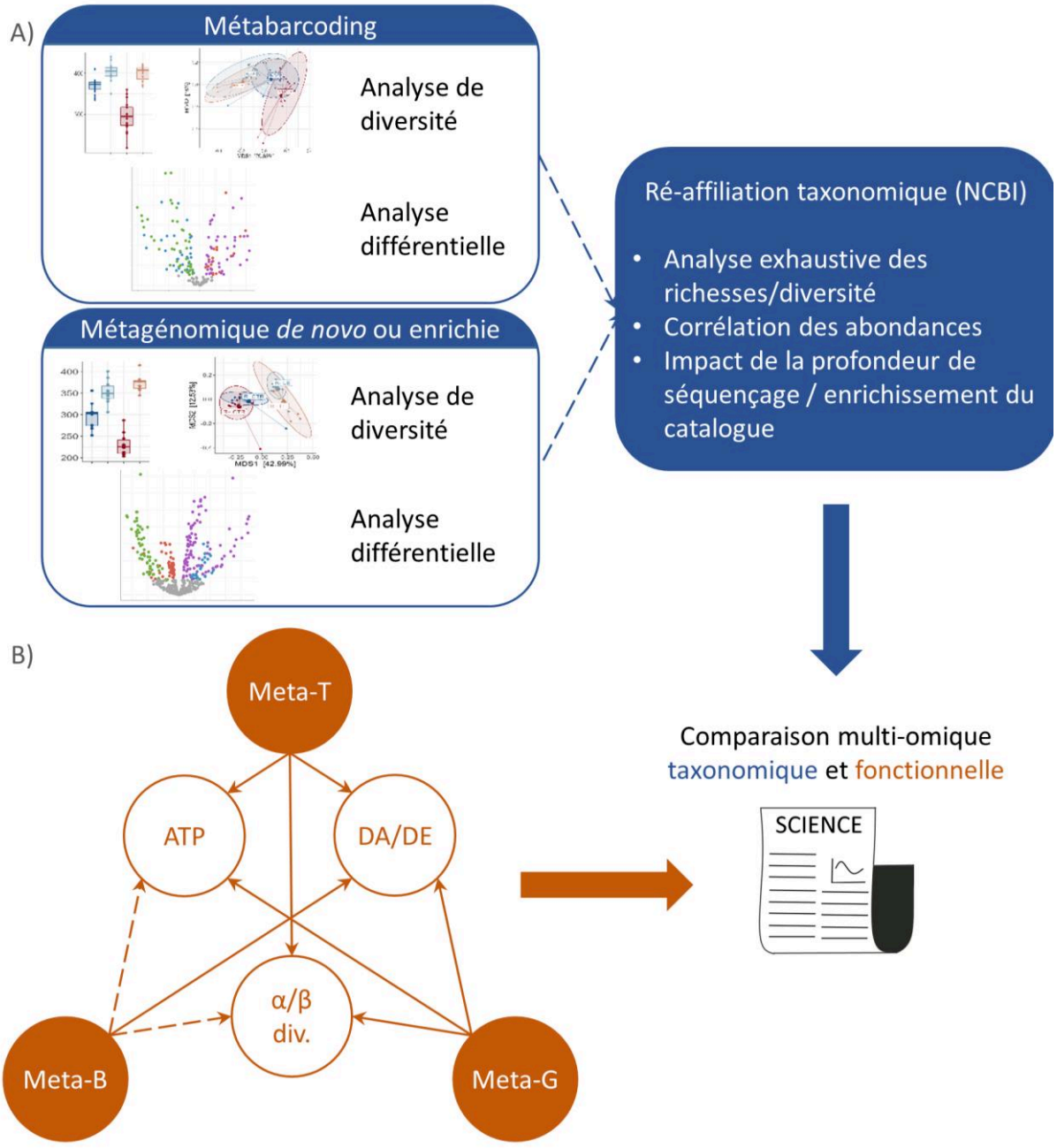


Figure G-2: Analyses comparatives multi-omiques.
A) à l'échelle des taxonomies identifiées par métabarcoding ou métagénomique ; B) à l'échelle des fonctions KEGG identifiées par métabarcoding (meta-B), métagénomique (meta-G) et métatranscriptomique (meta-T). ATP : Analyse Triadique Partielle ; α/β div. : Analyse de diversité alpha et bêta ; DA/DE : Analyse différentielle d'abondance/d'expression. Les flèches pointillées indiquent les analyses à compléter.

Ainsi, alors que plusieurs études ont comparé les profils taxonomiques obtenus à partir des données de métabarcoding et de métagénomique (Tessler et al. 2017; Knight et al. 2018; Brumfield et al. 2020; Durazzi et al. 2021), et que d'autres ont comparé les profils fonctionnels obtenus à partir de données de métagénomique et de métatranscriptomique (Franzosa et al. 2014; Mehta et al. 2018; F. Li et al. 2019), il n'y a pas à notre connaissance d'études comparant ces trois types de données omiques sur un écosystème complexe comme celui du microbiote cæcal de la poule. Cette étude comparative fera donc l'objet d'une seconde publication.

G.3. Tenir compte d'*a priori* pour l'analyse des métatranscriptomes

Pour cette première analyse des fonctions exprimées par le microbiote, nous avons choisi de travailler sans *a priori*, *i.e.* de travailler sur l'ensemble des annotations fonctionnelles et à partir des mesures d'expression uniquement. Or notre analyse comparative des profils d'abondance (données de métagénomique) et d'expression (données de métatranscriptomique), nous indique que les expressions sont fortement corrélées à leur abondance. Cette corrélation s'observe également au niveau des catégories fonctionnelles impactées par la lignée et/ou le régime. En effet, un certain nombre de fonctions ne représentent pas des activités métaboliques à proprement parler mais des fonctions de bases impliquées dans le fonctionnement de la cellule (par exemple le traitement de l'information génétique) ou dans la formation de composants structurels (par exemple la membrane). Or ces fonctions diffèrent selon le type de micro-organismes et *in fine*, ces catégories fonctionnelles reflètent les différences d'abondances de certains micro-organismes (par exemple les archées, ou les bactéries Gram négatives ou positives).

Pour compléter cette première exploration des métatranscriptomes, nous prévoyons d'ajouter deux nouvelles analyses :

- La première analyse consistera à cibler un métabolisme d'intérêt plutôt que de prendre l'intégralité des fonctions. Celui des carbohydrates nous semblent particulièrement pertinent. En effet, une des activités du microbiote consiste à dégrader les carbohydrates non digérés par les poules elles-mêmes et cette dégradation aboutit notamment à la production de SCFA qui sont connus pour directement impacter le métabolisme de l'hôte. Or l'annotation fonctionnelle de notre catalogue enrichi de gènes semble sous-estimer le nombre de CAZymes (ces enzymes dédiées à la synthèse et la dégradation des carbohydrate) (paragraphe **E.3.2**). Nous procéderons donc à une annotation complémentaire des gènes et des MGS en utilisant des outils spécialisés (dbCAN3 (Zheng, Ge, et al. 2023), dbCAN-seq (Zheng, Hu, et al. 2023)). Ces outils permettent outre

la détection et l'annotation des CAZymes, d'estimer le type de substrat (type de carbohydrate), et d'identifier les transporteurs et facteurs de transcription auxquels elles sont associées (**Figure G-3**).

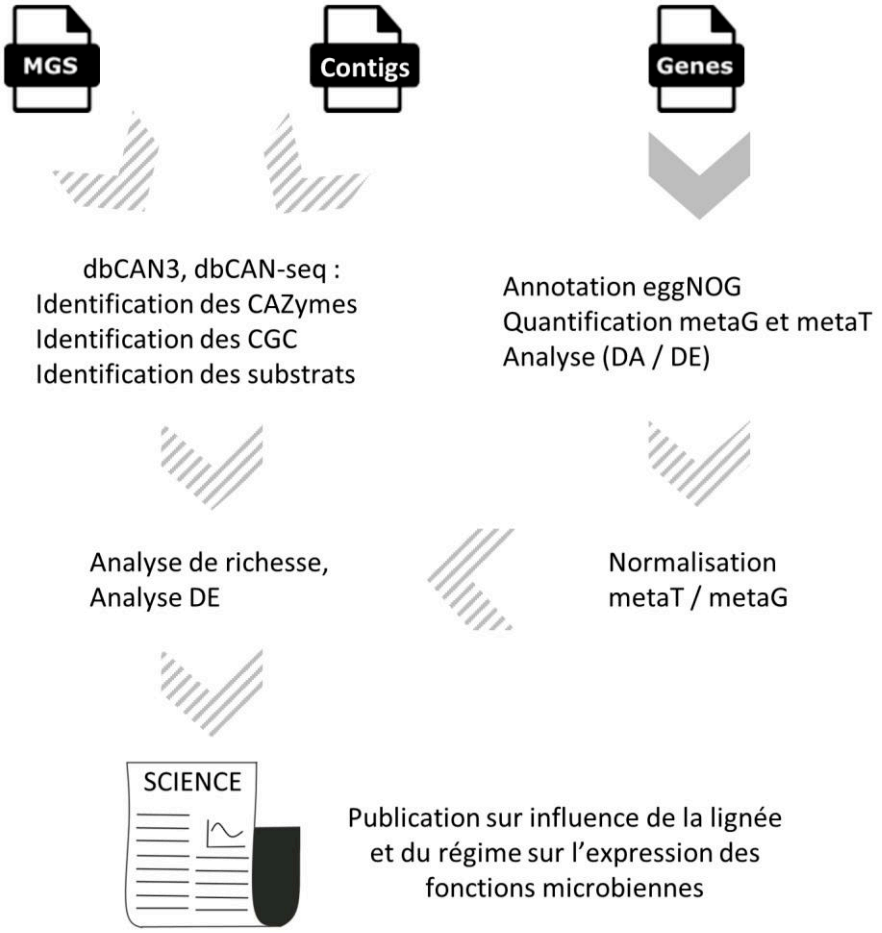


Figure G-3 : Processus d'analyses des métatranscriptomes.
Le chevron gris plein indique les analyses réalisées, les chevrons hachurés indiquent les analyses complémentaires à réaliser. Analyse DA / DE : Analyse différentielle d'abondance ou d'expression ; metaG : métagénomique ; metaT : metatranscriptomique.

- La seconde analyse consistera à tenir compte des abondances des fonctions pour normaliser leur expression. En effet, en ne travaillant que sur les expressions, comme nous l'avons observé, il est difficile de différencier une différence de transcription due à une différence d'abondance des fonctions (reflétant une différence d'abondance des micro-organismes), d'une différence due à une réelle activation ou répression de l'activité transcriptionnelle. L'une des façons de faire est de travailler sur les ratios entre expression et abondance (Yancong Zhang et al. 2021; Ji and Ma 2023). En appliquant cette normalisation des expressions nous espérons réduire le nombre de catégories fonctionnelles qui ne seraient qu'exclusivement due à une différence d'abondance et que cela mette en évidence des métabolismes plus pertinents pour discriminer nos conditions (**Figure G-3**).

A noter que ces dernières années, plusieurs études en métatranscriptomique ont comparé les outils initialement dédiés à la transcriptomique, les méthodes de normalisation des comptages, ainsi que la prise en compte ou non des abondances des fonctions et des taxonomies (Klingenberg and Meinicke 2017; Yancong Zhang et al. 2021; Ji and Ma 2023; Cho et al. 2023). On peut supposer que dans les années à venir des méthodes dédiées à la métatranscriptomique se développeront et que les stratégies d'analyses se stabiliseront pour identifier avec plus de robustesse les gènes et fonctions différentiellement exprimés.

Avec ces analyses complémentaires, nous prévoyons de valoriser ces données de métatranscriptomique dans un troisième article scientifique.

Par ailleurs, les travaux d'annotation fonctionnelle du catalogue MetaChick *via* eggNOG, ou prochainement KofamScan ou des CAZymes pourraient contribuer à valoriser cette ressource dans un quatrième article.

H. DISCUSSION

Dans cette étude, nous avons cherché à évaluer l'impact sur la composition du microbiote d'une sélection divergente de longue durée effectuée sur la consommation alimentaire résiduelle. L'approche multi-omique mise en place pour cela a permis d'étudier l'association possible entre la composition du microbiote cæcal et l'efficience alimentaire des poules pondeuses. En effet, il est bien établi que l'efficience alimentaire, phénotype étudié depuis de nombreuse années chez différentes espèces d'animaux d'élevage, est sous le contrôle de facteurs génétiques et environnementaux. Mais en outre, les multiples interactions entre le microbiote intestinal et son hôte, en particulier grâce à ses capacités de dégradation des aliments non digestes par l'hôte lui-même, font du microbiote cæcal un facteur qui pourrait également influencer l'efficience alimentaire. Etant donné que ces deux éléments (efficience et microbiote) sont influencés par la composition du régime alimentaire, nous avons basé cette thèse sur un dispositif expérimental impliquant deux lignées expérimentales de poules pondeuses divergentes pour l'efficience alimentaire (R+ et R-) nourries avec deux régimes alimentaires se différenciant sur leur contenu en amidon et en fibre (CTR et LE).

D'un point de vue méthodologique, il existe diverses approches de caractérisation des écosystèmes microbiens. Ces approches se différencient par la question de recherche à laquelle elles vont pouvoir répondre, et sur l'exhaustivité de la caractérisation des communautés qui seront prises en compte. Lors de cette étude, nous avons utilisé trois méthodes permettant d'étudier le microbiote cæcal dans son ensemble (*i.e.* sans culture préalable des micro-organismes), et de le caractériser du point de vue des communautés microbiennes présentes (*via* leur taxonomie) et du point de vue fonctionnel (*via* leurs gènes) : le métabarcoding, la métagénomique et la métatranscriptomique. Ces méthodes sont à la fois complémentaires et partiellement chevauchantes dans leurs objectifs de recherche et présentent également des difficultés différentes de protocole expérimental et d'analyse.

Résumé des résultats principaux et perspectives

- **D. Partie 1 : L'analyse de données de métabarcoding révèle des interactions hôtes-microbiotes dépendantes du régime**

Données : métabarcoding
Résultats :
- La sélection divergente a abouti à une différenciation du microbiote.
- Le microbiote pourrait contribuer à l'efficience alimentaire *via* une production différenciées des SCFA.
- Ces associations sont dépendantes du régime alimentaire.

Perspectives :
- Préciser les affiliation taxonomiques (*via* la métagénomique).
- Valider l'activité des fonctions microbiennes d'intérêt (*via* la métatranscriptomique).

- **E. Partie 2 : Constitution d'un catalogue enrichi de taxonomies et de fonctions.**

Données : métagénomique
Résultats :
- L'analyse de novo par co-assemblage a permis la prise en compte de profondeurs variables de séquençage.
- La combinaison de l'analyse *de novo* et du catalogue pré-existant (MetaChick) permet de couvrir une large diversité microbienne.
- Les effets de la lignée et du régime sur les indices de richesse et de diversité sont confirmés à l'échelle des taxonomies, mais très partiellement à l'échelle des fonctions.
- La précision des affiliations taxonomiques est améliorée.

Perspectives :
- Comparer exhaustivement les taxonomies obtenues grâce à la métagénomique et au métabarcoding après réaffiliation des entités taxonomiques.

- **F. Partie 3 : Analyse de données de métatranscriptomique : différenciation fonctionnelle des microbiotes cæcaux.**

Données : métagénomique et métatranscriptomique
Résultats :
- Les indices de diversité montrent que la variabilité des fonctions exprimées (*metaT*) selon les individus, la lignée et le régime est plus grande que celle des fonctions présentes (*metaG*), mais reste moindre que celle observée entre les espèces (MGS, *metaG*).
- Il existe une corrélation importante entre abondance et expression des fonctions.
- Les fonctions différentiellement exprimées représentent en priorité des différences de taxonomies.

Perspectives :
- Comparer exhaustivement les fonctions identifiées grâce aux trois omiques.
- Analyser spécifiquement les CAZymes.
- Normaliser les expressions des fonctions par leurs abondances.

H.1. Le rôle du microbiote cæcal dans l'efficience alimentaire des poules pondeuses : origine de sa différenciation

Dans un contexte de nutrition optimale (régime CTR), l'une des caractéristiques clés de la différenciation des microbiotes associés à l'une ou l'autre des lignées est cette très grande différence de richesse microbienne : « the richer the better ». Notre hypothèse est que cette plus grande richesse chez la lignée efficiente R-, lui apporte une capacité plus importante à exploiter les ressources nutritives, ce qui peux effectivement représenter un atout pour augmenter l'efficience. Par ailleurs, les variations de composition des communautés microbiennes et des fonctions associées nous ont amené à démontrer une différence de production d'acide gras à chaine courte (SCFA), métabolites particulièrement importants pour le métabolisme de l'hôte.

Ces différences de composition du microbiote entre lignées pourraient être attribuées à plusieurs facteurs. Une première hypothèse serait que l'environnement ait contribué à influencer la composition du microbiote. En effet, celui-ci est connu pour être un ensemble de facteurs qui influence particulièrement le microbiote (Rothschild et al. 2018; Richards et al. 2019; Tabrett and Horton 2020). Or l'environnement a été contrôlé au maximum tout au long de l'expérimentation pour qu'il soit identique pour les deux lignées. Ainsi, les œufs dont la fécondation a eu lieu au même moment, ont été placés dans les mêmes incubateurs et éclosoirs, ont éclos le même jour et les poussins des deux lignées ont été élevés ensemble jusqu'à l'âge adulte. On peut donc supposer que dans notre contexte, l'environnement n'a pas eu d'effet significatif pour expliquer les différences observées entre les microbiotes des deux lignées.

Une seconde hypothèse serait que les différences de composition du microbiote soient imputables à un effet de la génétique de l'hôte. En effet, il est désormais admis que celle-ci peut influencer la composition du microbiote intestinal (Goodrich et al. 2016; Mittal et al. 2019; Cahana and Iraqi 2020; Fan et al. 2020; Wen et al. 2021). Il a par ailleurs été démontré que la composition du microbiote est héritable, par exemple chez le poulet de chair (Mignon-Grasteau et al. 2015), ou qu'elle pouvait être utilisée en sélection par exemple chez le porc (Larzul et al. 2024). Dans notre contexte, les deux lignées descendent de la même population parentale et ont divergé phénotypiquement et génétiquement au cours des générations sous l'effet de la sélection sur la RFI, caractère qui a démontré être héritable (Tixier-Boichard et al. 1995; Jehl 2020). Ainsi, en considérant que la génétique de l'hôte est en partie à l'origine de son efficience alimentaire et qu'elle est aussi supposée pouvoir moduler la composition de son microbiote, nous pouvons émettre l'hypothèse que les associations identifiées entre l'efficience alimentaire et le microbiote sont le fruit des différences génétiques entre lignées. Pour valider cette hypothèse et identifier les gènes influençant la composition du microbiote une approche serait de

réaliser des analyses d'association (GWAS). Toutefois cela nécessiterait de produire un plus grand nombre d'animaux avec des données de génotypage et bien sûr de composition microbienne.

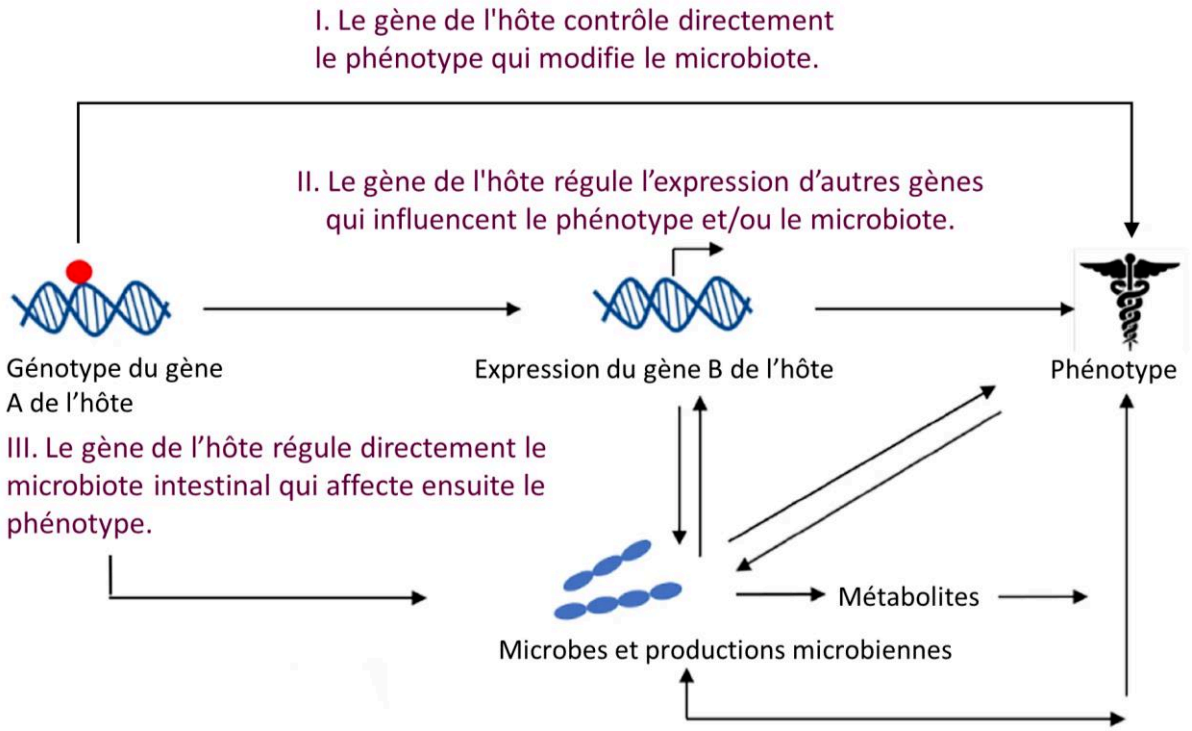


Figure H-1 : Modèle représentant les voies directes et indirectes de l'influence de la génétique sur le microbiote et un phénotype.
Adaptée de Bubier, Chesler, and Weinstock 2021

Les mécanismes d'action de la génétique peuvent être multiples. Ils peuvent agir directement ou bien indirectement sur la composition du microbiote (**Figure H-1**) (Bubier, Chesler, and Weinstock 2021). Une action directe pourrait consister en la modulation de l'activité de gènes spécifiques affectant directement l'abondance de certains microbes. Une action génétique indirecte pourrait impliquer la modification d'un phénotypes qui altèrerait ensuite la composition microbienne. Bien que notre dispositif ne permette pas d'aller plus loin dans l'identification des mécanismes d'influence de la génétique sur la composition du microbiote, il est probable que plusieurs mécanismes existent, avec des modes d'action à la fois directs et indirects. Par exemple, le système immunitaire est un élément clef dans le maintien de l'homéostasie intestinale et l'inhibition de l'inflammation. Grâce à la production de peptides antimicrobiens, de mucines ou d'immunoglobulines, il influence directement la prolifération de certains micro-organismes (Robinson et al. 2015; Khasanah et al. 2024). Par ailleurs, cette relation système immunitaire – microbiote est une relation bidirectionnelle, le microbiote pouvant également influencer la production de ces composants. Or, de façon intéressante, nos deux

lignées ont montré des réponses immunitaires différentes au niveau des systèmes immunitaires innés et adaptatifs (Zerjal et al. 2021). La prise alimentaire, quant à elle, pourrait être un exemple de phénotype intermédiaire de l'action indirecte de la génétique sur la composition du microbiote. En effet, la sélection sur la RFI a également abouti à une divergence forte sur la quantité d'aliment ingéré (« Feed Intake », FI). En outre, la quantité ingérée d'un composant particulier est un facteur connu pour influencer sa digestibilité par l'hôte ou celle d'autres composants (Svihus 2011; Tejeda and Kim 2021). Pour aller plus loin dans la caractérisation des lignées R+ et R-, il pourrait être intéressant d'évaluer leurs capacités propres à digérer différents composants alimentaires, en particulier l'amidon. Il a été aussi prouvé que la FI, indépendamment de la RFI, influence des caractéristiques physiques des organes intestinaux, la composition microbienne de différents segments intestinaux et des fèces, ainsi que la production d'acides gras au niveau du cæcum (Siegerstetter et al. 2018; Metzler-Zebeli et al. 2019). Dans notre étude, la corrélation phénotypique entre la RFI et la FI est de 0,95, indiquant que ces deux caractères sont intrinsèquement liés et ne peuvent être différenciés. Si nous considérons que la différence microbienne est la résultante d'une différence quantitative de prise alimentaire entre les lignées, nous pourrions nous attendre à obtenir le même résultat quel que soit le type de régime. Or, ce n'est pas le cas.

En effet, dans la condition de régime appauvri en énergie (LE), les différences entre lignées à l'échelle du microbiote (compositions taxonomique et fonctionnelle) sont presque inexistantes alors que les différences de RFI et de FI sont maintenues. Ces résultats illustrent la forte interaction entre la génétique et la composition du régime sur la composition du microbiote et probablement sur la contribution de ce dernier à l'efficience alimentaire des poules. On peut supposer que les années de sélection réalisées avec le régime alimentaire CTR, ont abouti à la mise en place d'une différentiation du microbiote directement liée à la composition de ce régime.

H.2. Importance de l'impact du régime : défis et opportunités

Dans notre étude, la modification du régime alimentaire est le facteur principal de structuration du microbiote. La composition microbienne (taxonomies) des deux lignées est largement impactée, avec notamment une augmentation de la richesse dans la condition de régime LE. Bien que les fonctions portées et exprimées par le microbiote soient moins sensibles que les taxonomies aux différents facteurs génétiques et environnementaux, elles présentent également des variations significatives suite à la modification du régime alimentaire. Cette forte influence du régime sur le microbiote, bien que documentée chez différentes espèces animales (York 2019; Kogut 2022), n'est pas toujours bien explicitée ou argumentée dans les études sur le lien entre le microbiote et un phénotype de l'hôte.

L'influence du régime alimentaire est telle que la comparaison bibliographique doit, en plus d'être réalisée autant que possible sur la même espèce hôte, être contextualisée au régime alimentaire utilisé. Cependant, le régime alimentaire n'est pas toujours indiqué ou décrit avec la même précision, ce qui complexifie la comparaison des études et pourrait expliquer une partie des contradictions de la littérature. Par exemple, Yan et al. 2017 ont analysé le microbiote cæcal de poule pondeuse adulte et ont identifié une association entre l'abondance du genre *Lactobacillus* et une meilleure efficience alimentaire (mesurée grâce à la RFI). Nos deux études se positionnent *a* priori dans un contexte expérimental similaire. Pourtant nos résultats se contredisent puisque dans nous avons observé que ce genre tend à être associé à une faible efficience. De nombreux facteurs pourraient expliquer cette contradiction : l'âge des animaux (60 semaines contre 31 semaines), la souche (pondeuse naine à œuf brun contre Rhode Island Red), l'environnement d'expérimentation et la méthode d'analyse, mais le régime alimentaire pourrait en particulier en être responsable. En effet, nos résultats indiquent une association importante de *Lactobacillus* avec le régime CTR en comparaison au régime LE or l'étude de Yan *et al.* ne précise pas la composition du régime utilisé.

L'utilisation de nos deux régimes nous a permis de mettre en évidence l'impact de la quantité d'aliments ingérés et de la composition générale du régime, mais également d'émettre l'hypothèse que la richesse des variétés de céréales impactait particulièrement la composition du microbiote. En effet, le régime LE contient une diversité de céréales et donc de fibres plus importante que dans le régime CTR et chacune de ces fibres peut être dégradée par un nombre restreint voire spécifiques de micro-organismes (Cantu-Jungles and Hamaker 2023). Ces micro-organismes auront à leur tour des effets plus ou moins marqués sur le métabolisme du microbiote ou de l'hôte. De fait, il n'est pas surprenant que de nombreuses études s'intéressent précisément à l'impact de certains composants alimentaires (en particulier les céréales contenant des fibres indigestes par l'hôte) pour tenter de moduler le microbiote vers une composition bénéfique pour l'hôte. Ces composants correspondent à ce que l'on appelle des prébiotiques. Des revues récentes dressent un bilan des effets observés à ce jour de l'utilisation des prébiotiques chez les poules (Khan et al. 2020; Leone and Ferrante 2023). Ainsi différents prébiotiques dérivés de fibres végétales sont associées à plusieurs caractères de croissances, d'efficience alimentaire ou de santé en jouant directement sur la digestibilité des nutriments, les caractéristiques physiques de l'intestin et la composition microbienne. Inversement d'autres études cherchent à identifier les micro-organismes (probiotiques) ou métabolites naturellement produits par le microbiote (postbiotiques) qui permettent notamment une meilleure dégradation des aliments et

in fine de meilleures performances (Jansseune et al. 2024). Les recherches doivent se poursuivre pour définir le bon équilibre entre la composition du régime alimentaire, ces additifs, et les bénéfices pour l'hôte.

H.3. Vers une description exhaustive des communautés, des fonctions et de leur association

Du point de vue méthodologique, les trois méthodes omiques utilisées dans cette thèse ont un premier objectif théorique de description exhaustive des écosystèmes microbiens à deux échelles (taxonomique et/ou fonctionnelle). Or chacune de ces méthodes réalise cet objectif avec un succès variable et dépendant du type d'annotation, taxonomique ou fonctionnelle.

H.3.1. Focus vers l'exhaustivité taxonomique

L'identification des taxonomies présentes dépend de plusieurs facteurs méthodologiques influencés par des facteurs biologiques. Pour les données de métabarcoding, l'exhaustivité des taxonomies quantifiées est fortement influencée par le choix du gène marqueur. Pour les données de métagénomique, théoriquement, si les lectures sont produites en suffisamment grande quantité, elles ne doivent pas souffrir de ce biais de représentativité. Toutefois, la capacité à identifier chaque génome dépendra de la qualité de l'assemblage des lectures en contigs puis en MGS. Or, ces traitements sont influencés par l'abondance des micro-organismes et *in fine* la représentativité des taxonomies sera biaisée vers les génomes les plus abondants. Par ailleurs, nous avons pu voir que globalement les données de séquençage sont fortement bruitées et qu'il est donc nécessaire de filtrer les entités taxonomiques (ASV, MGS) selon des seuils d'abondance et de prévalence. De fait, l'analyse de la richesse rare d'un écosystème microbien est difficilement appréhendable, et ceci est d'autant plus vrai pour des écosystèmes qui présentent une richesse importante et des abondances fortement déséquilibrées, comme les microbiotes intestinaux.

Malgré ces constations, d'autres stratégies d'utilisation et des développements méthodologiques permettent d'améliorer le nombre d'organismes identifiés. A l'image de ce qui est fait dans les études sur les ADN environnementaux dont le but est le suivi de la richesse globale de l'environnement, l'une des stratégies pourrait être d'utiliser un ensemble de gènes marqueurs complémentaires permettant de couvrir au mieux l'ensemble de la richesse taxonomique (Brandt et al. 2021; Porter and Hajibabaei 2022). L'inconvénient de cette technique est la difficulté d'intégration des résultats obtenus à partir de chaque marqueur analysé indépendamment des uns et des autres (Da Silva et al. 2019). Pour la

métagénomique, l'analyse de la richesse microbienne peut se réaliser à une autre échelle que celle des MGS. Il existe en particulier des outils qui utilisent directement les lectures, soit pour une affiliation taxonomique directe (comme Kraken2 (Wood, Lu, and Langmead 2019)), soit pour la quantification de gènes marqueurs de chaque taxon (comme MetaPhlAn3 (Blanco-Miguez et al. 2022)). Bien que ces stratégies évitent l'écueil de l'assemblage des métagénomes, elles ne reposent que sur la quantification de génomes connus et leurs résultats sont donc particulièrement influencés par la base de données sur laquelle elles se basent (Wright, Comeau, and Langille 2023). Pour notre étude, nous avons pris le parti de combiner notre analyse *de novo* avec un catalogue de MGS dédié à notre écosystème microbien, et nous avons proposé dans les chapitres précédents des réflexions qui pourraient permettre d'améliorer encore la richesse détectée (paragraphe **G.1**).

Une solution qui semble prometteuse à la fois pour le métabarcoding et pour la métagénomique est l'utilisation de séquenceurs permettant la génération de longues lectures (PacBio, ONT). En métagénomique, ces lectures réduisent fortement la fragmentation des assemblages, aboutissant même à l'obtention de génomes complets et circulaires, et pourraient permettre la détection de nouveaux organismes comparés aux génomes assemblés à partir de lectures courtes (Yan Zhang et al. 2022; Eisenhofer et al. 2024). Parallèlement, dans le cadre des analyses de métabarcoding, ces technologies « longues-lectures », permettent de couvrir l'ensemble du gène de l'ARNr 16S, voire l'ensemble de l'opéron 16S-ITS-23S, ce qui améliore drastiquement la discrimination des espèces procaryotes, ainsi que la capacité à identifier précisément la taxonomie (Santos et al. 2020; Mainguy et al. 2022; Gulyás et al. 2023; Buetas et al. 2024). Elles nécessitent toutefois des mises au point de protocoles expérimentaux et analytiques (Mainguy et al. 2022; Gulyás et al. 2023). En effet, comparée à la technologie Illumina réputée la plus fiable avec un fort débit, les profils d'erreurs sont différents (particulièrement pour ONT), et le taux d'erreurs a longtemps été bien supérieur, même si cela tend à ne plus être significatif aujourd'hui. Par ailleurs, le débit et le multiplexage est plus faible pour la technologie Sequel II de PacBio augmentant ainsi les coûts. D'un point de vue analytique, identifier une base de référence permettant l'assignation taxonomique de ces séquences longues, telle que la région 16S-ITS-23S, reste primordial. A ce jour, on peut noter quelques initiatives très récentes de construction de bases de données de richesse bactérienne variables (Seol et al. 2022; Pascal 2023; Walsh et al. 2024).

Finalement le choix de la base de référence est le dernier point clé pour caractériser les entités taxonomiques détectées (ASV, MGS). Les résultats varieront particulièrement d'une part en fonction de sa complétude vis-à-vis de l'écosystème étudié et d'autre part en fonction de la qualité de sa classification taxonomique. Ces dernières années, grâce aux nombreuses études utilisant notamment des techniques de séquençage à haut débit sur une variété grandissante d'écosystèmes plus ou moins

complexes, les bases de données de référence se sont enrichies d'un nombre colossal de séquences annotées taxonomiquement. Dans le cas des organismes procaryotes, on peut citer l'exemple de la base GTDB qui, sur les cinq dernières années, a plus que quadruplé son nombre de séquences de génomes (dont des MGS) et d'espèces qui s'élève désormais à plus de 113 000. A noter que le nombre de génomes de haute qualité a doublé sur cette période mais celui des génomes de moyenne qualité a été multiplié par six. Par ailleurs, l'ajout continu de nouvelles séquences fait évoluer les classifications taxonomiques. Ainsi pour GTDB, par rapport à la taxonomie du NCBI, 65% des génomes ont vu leur taxonomie corrigée ou complétée (soit +20% en cinq ans). En revanche, les observations sont différentes concernant la base de données SILVA. On constate que le nombre de séquences de bonne qualité évolue plus lentement, +6% en trois ans, dépassant ainsi les 2 millions de séquences du gène de l'ARNr 16S, mais le nombre d'espèces identifiées sur les séquences non redondantes a diminué de 14% (désormais à 42 996). Ceci est le fait d'une modification de la méthode employée pour éliminer la redondance.

Avec l'évolution des connaissances dans ces bases de référence, on peut imaginer que la caractérisation de la richesse taxonomique s'en retrouve améliorée notamment pour des écosystèmes historiquement moins étudiés. Cela indique également l'importance d'utiliser des références spécialisées les plus récentes possibles qui affinent, complètent ou corrigent les annotations taxonomiques des bases de référence généralistes.

H.3.2. De l'annotation fonctionnelle à l'identification taxonomique des contributeurs

Alors que l'annotation taxonomique évolue rapidement, il n'en est pas de même pour l'annotation fonctionnelle. En effet, même sur des écosystèmes largement étudiés comme celui du microbiote intestinal humain, selon la méthode d'annotation, entre 40 et 70% des gènes restent sans fonction connue (Heintz-Buschart and Wilmes 2018), et la précision des annotations varient selon la base de référence utilisée. Une des explications est le développement méthodologique à plusieurs vitesses selon les domaines d'études. En effet, la caractérisation des écosystèmes microbiens grâce à l'utilisation de protocole de séquençage a permis de lever la limite de l'analyse des micro-organismes cultivables. Cependant, ils ont également incité les scientifiques à générer des volumes massifs de données, un processus devenu de moins en moins complexe et de plus en plus abordable financièrement. Pour répondre à cet afflux de données, la priorité a été mise sur le développement d'outils bioinformatiques permettant notamment le traitement automatique des séquences, au détriment d'outil et de personnel dédiés à l'annotation et la curation manuelle de celles-ci (Vincent et al. 2017). L'annotation fonctionnelle nécessite la mise en place d'expérimentations dédiées, à l'échelle

de chaque organisme, d'un gène ou d'une protéine. Ces analyses se font donc sur un temps long et un débit plus modéré, un peu à contre-courant des analyses basées sur le séquençage, creusant ainsi un fossé entre la caractérisation exhaustives des organismes et la caractérisation fine des fonctions des génomes (Berg et al. 2020). Toutefois, le nombre de génomes annotés fonctionnellement pris en compte dans les bases de données d'annotations fonctionnelles progresse. La base de données eggNOG 6.0 inclut 12 535 espèces (+7 445 en 3 ans), tandis que la base de données KEGG contenait 8 400 génomes en 2022 et en contient actuellement 10 247, représentant 7 680 espèces. A l'image des analyses taxonomiques, les annotations évoluent fortement et il convient d'utiliser autant que possible les dernières versions de ces bases pour éviter de sous-estimer le potentiel fonctionnel de l'écosystème étudié (Wadi et al. 2016).

Au-delà de l'identification la plus exhaustive possible des fonctions portées par le microbiote, c'est l'identification de celles qui varient entre conditions que l'on cherche à établir. Les méthodes pour réaliser ce type d'analyses sont nombreuses (et très discutées) aboutissant à des résultats différents (Manor and Borenstein 2017; Yancong Zhang et al. 2021; Douglas and Langille 2021; Ji and Ma 2023). Avec quelle entité fonctionnelle travailler ? Comment normaliser les abondances ? Faut-il tenir compte des annotations taxonomiques, des prévalences dans différentes espèces ou des abondances fonctionnelles dans le cas de la métatranscriptomique ?

En particulier pour le choix de l'entité fonctionnelle, le gène est le grain le plus fin porteur de l'information fonctionnelle. Mais, dans le cadre d'analyse globale des fonctions qui varient entre deux conditions et comme cela est généralement fait, nous avons choisi d'analyser les abondances à l'échelle de familles de gènes annotées. Cela permet de gérer la forte parcimonie des données d'abondance due à la forte individualisation de la composition du microbiote (Heintz-Buschart and Wilmes 2018; Douglas and Langille 2021). Certaines études agglomèrent les abondances des gènes à une granularité plus importante encore, par exemple à la voie métabolique, ce qui permet une interprétation plus facile. Toutefois cela peut aussi mener à de fausses interprétations si les fonctions centrales ne sont pas présentes. De plus, à cette échelle, on perd également en précision mécanistique. Bien que travailler à l'échelle des familles de gènes ait ses avantages, cela implique la déconnection entre la fonction et la taxonomie, ce qui est le défaut de la majorité des études métagénomiques (Douglas and Langille 2021). Or cette relation est d'un intérêt primordial si l'objectif à terme est d'agir sur la composition d'un microbiote afin d'activer ou de réprimer une fonction. Sur les données de métagénomique, chaque gène peut être identifié taxonomiquement *via* les bases généralistes et la taxonomie du NCBI en particulier, mais ces annotations taxonomiques ne sont pas très précises ou complètes et comme indiqué précédemment ne sont pas nécessairement cohérentes avec celles de la banque GTDB généralement utilisée pour l'annotation des MGS. Une autre solution serait d'aligner

directement les gènes sur les MGS, mais une part d'entre eux ne seraient pas alignés et l'information resterait également incomplète.

Malgré ces manques ou défaut d'interopérabilité, les nombreuses études qui explorent de manière globale les écosystèmes microbiens enrichissent les bases généralistes et spécialisées, et font évoluer les taxonomies et les orthologies. Par ailleurs, l'analyse des abondances des entités taxonomiques (ASV, MGS), ou fonctionnelle (gènes, familles de gènes) dans différentes conditions permet de détecter les taxonomies, ou métabolismes d'un intérêt plus particulier. C'est en partie le cas pour la poursuite de notre étude durant laquelle nous prévoyons l'analyse détaillée de l'activité des carbohydrates par le microbiote cæcal de poule pondeuse. Par la suite, ces entités pourront représenter les cibles d'une recherche de précision, aboutissant en particulier à l'amélioration des annotations taxonomiques et fonctionnelles, ainsi que de leur association.

I. CONCLUSION ET BILAN PERSONNEL

Cette thèse représente un projet ambitieux sur les plans à la fois scientifique et personnel.

Sur le plan scientifique, elle visait d'une part à identifier les acteurs microbiens influençant l'efficience alimentaire des poules pondeuses dans différents contextes de régime alimentaire et à comprendre leurs mécanismes d'action, et, d'autre part, à comparer et intégrer trois méthodes de séquençage omiques, développées plus ou moins récemment, chacune présentant un degré variable de maîtrise et de standardisation dans le traitement des données.

L'analyse des données de métabarcoding a permis d'émettre plusieurs hypothèses cohérentes avec la littérature et suggérant de nouvelles analyses pour les préciser. Elle a également illustré les limites déjà connues de cette technique auxquelles les données de métagénomique semblent pouvoir répondre. Ces dernières données, bien que de plus en plus utilisées par la communauté scientifique, ont nécessité une mise au point du processus d'analyse importante, et les analyses doivent se poursuivre pour les exploiter entièrement et les valoriser. Enfin les données de métatranscriptomique ont montré une réelle plus-value comparée aux deux précédentes. En effet, bien que l'expression des gènes (et par extension des fonctions) soit fortement corrélée à leur abondance, les métatranscriptomes ont montré une plus grande variabilité aux facteurs génétiques et environnementaux. Par ailleurs, un nombre important de fonctions abondantes n'ont pas été détectées comme actives. Toutefois, la mise au point expérimentale ainsi que du processus d'analyse demande encore des développements, et à l'image des données de métagénomique, celles-ci ne seront valorisées qu'après des analyses complémentaires. Pour conclure sur la comparaison qualitative de l'utilisation de ces trois techniques omiques, elles représentent à elles trois un véritable gradient liant précision, exhaustivité et coût d'un côté et facilité expérimentale, d'analyse et d'interprétation de l'autre. A mon sens le métabarcoding reste une solution tout à fait pertinente pour une première exploration d'un écosystème microbien par un scientifique ayant peu de connaissances *a priori* des protocoles expérimentaux et des méthodes d'analyses. D'un autre côté, la métagénomique et la métatranscriptomique représentent des outils de pointe (bien que produisant des données en masse) qui nécessitent des connaissances et des compétences importantes et dont l'interprétation est facilitée si on les aborde avec des *a priori*.

Sur le plan personnel, cette thèse a représenté un véritable défi. Après près de 15 ans d'activité en tant qu'ingénieure dans différentes structures d'accompagnement à l'analyse de séquences, j'ai débuté cette thèse avec de fortes connaissances générales sur le séquençage, la gestion des séquences, leur analyse bioinformatique, ainsi qu'en développement d'outils notamment à destination d'utilisateurs non informaticien, mais peu de connaissances en biologie et en analyse fonctionnelle. Ces dernières années j'ai pu me consacrer en particulier à l'analyse de données de métabarcoding *via* le développement de la suite FROGS. Mais finalement, je n'ai que rarement eu l'occasion de me confronter à l'analyse de données en vue de répondre à une problématique biologique. Plus précisément, la motivation personnelle de cette thèse a été de porter tout ou partie d'un projet scientifique, de la recherche de financement à la valorisation des résultats, de la mise en place d'expérimentations à celle des processus d'analyses bioinformatiques et biostatistiques en passant bien sûr par l'interprétation biologique des résultats. La richesse de ces trois années n'aura pas été que microbienne. Elles m'auront permis de renouer ou de créer de nouvelles collaboration avec une multitude de personnes, qui illustrent la diversité des compétences nécessaires à la réalisation de ce type de projet, complexe sur bien des aspects. Elles m'ont permis d'acquérir de nouvelles connaissances et compétences dans de nombreux domaines : i) sur les protocoles expérimentaux et la mise en place de contrôle qualité qui même s'ils sont un peu éloignés de mes activités quotidiennes, enrichissent la façon d'appréhender les données ; ii) sur les procédures d'analyse en particulier en statistiques et bien sûr sur les outils spécialisés de bioinformatique ; iii) sur les bases de référence taxonomiques et fonctionnelles ainsi que sur certains métabolismes ciblés qui m'ont donné l'occasion de collaborer avec des experts. Ce projet va se poursuivre sur des analyses complémentaires, fruits des interprétations et réflexions menées jusque-là. Il ouvre également la porte sur de nouveaux développements méthodologiques en bioinformatique, et il élargit ma façon d'imaginer le plan d'analyse des données de métagénomique (au sens large) en fonction des problématiques biologiques posées, et de la complexité de l'écosystème étudié.

Charlie (10 ans), Tya (13 ans), Xavier (50 ans)

ANNEXES

J.1. Autres communications dans lesquelles je suis co-auteure

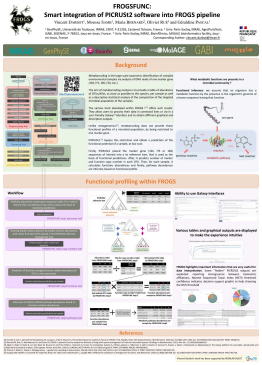
Vous trouverez en annexes les travaux valorisés par des communications dans lesquels je suis co-auteure :

- Une revue dans Molecular Ecology Resources qui répertorie les suites d'outils permettant l'analyse de séquences de métabarcoding et pour laquelle j'ai contribué en présentant la suite d'outils FROGS.

Hakimzadeh, A., Abdala Asbun, A., Albanese, D., **Bernard, M.**, Buchner, D., Callahan, B., Caporaso, J. G., Curd, E., Djemiel, C., Brandström Durling, M., Elbrecht, V., Gold, Z., Gweon, H. S., Hajibabaei, M., Hildebrand, F., Mikryukov, V., Normandeau, E., Özkurt, E., M. Palmer, J. … Anslan, S. (2024). A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses. *Molecular Ecology Resources*, 24, e13847. https://doi.org/10.1111/1755-0998.13847

- Un poster présenté lors de le conférence de bioinformatique JOBIM 2022 par Vincent Darbot, ingénieur en bioinformatique de l'unité GenPhySE que j'ai co-encadré sur la mise en place des outils d'inférence fonctionnelle dans FROGS.

Darbot, V., Samb, M., **Bernard, M.**, Rué, O., Pascal, G. FROGSFUNC: Smart integration of PICRUSt2 software into FROGS pipeline. *JOBIM 2022*, Jul 2022, Rennes, France. https://hal.inrae.fr/hal-03806133



Received: 10 February 2023 | Revised: 5 June 2023 | Accepted: 6 July 2023

DOI: 10.1111/1755-0998.13847

MOLECULAR ECOLOGY RESOURCES WILEY

FROM THE COVER

A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses

Ali Hakimzadeh[1] | Alejandro Abdala Asbun[2] | Davide Albanese[3] | Maria Bernard[4,5] | Dominik Buchner[6] | Benjamin Callahan[7] | J. Gregory Caporaso[8] | Emily Curd[9] | Christophe Djemiel[10] | Mikael Brandström Durling[11] | Vasco Elbrecht[6] | Zachary Gold[12] | Hyun S. Gweon[13,14] | Mehrdad Hajibabaei[15] | Falk Hildebrand[16,17] | Vladimir Mikryukov[1] | Eric Normandeau[18] | Ezgi Özkurt[16,17] | Jonathan M. Palmer[19] | Géraldine Pascal[20] | Teresita M. Porter[15] | Daniel Straub[21] | Martti Vasar[1] | Tomáš Větrovský[22] | Haris Zafeiropoulos[23] | Sten Anslan[1]

[1]Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia

[2]Department of Marine Microbiology and Biogeochemistry, NIOZ Royal Netherlands Institute for Sea Research, Texel, Netherlands

[3]Unit of Computational Biology, Research and Innovation Centre, Fondazione Edmund Mach, Italy

[4]INRAE, AgroParisTech, GABI, Université Paris-Saclay, Jouy-en-Josas, France

[5]INRAE, SIGENAE, Jouy-en-Josas, France

[6]Aquatic Ecosystem Research, University of Duisburg-Essen, Essen, Germany

[7]Department of Population Health and Pathobiology, College of Veterinary Medicine and Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, USA

[8]Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, Arizona, USA

[9]Vermont Biomedical Research Network, University of Vermont, Burlington, Vermont, USA

[10]Agroécologie, INRAE, Institut Agro, Univ. Bourgogne Franche-Comté, Dijon, France

[11]Department of Forest Mycology and Plant Pathology, Swedish University of Agricultural Sciences, Uppsala, Sweden

[12]Zachary Gold, NOAA Pacific Marine Environmental Laboratory, Seattle, Washington, USA

[13]UK Centre for Ecology & Hydrology, Oxfordshire, UK

[14]School of Biological Sciences, University of Reading, Reading, UK

[15]Department of Integrative Biology and Centre for Biodiversity Genomics, University of Guelph, Guelph, Ontario, Canada

[16]Gut Microbes & Health, Quadram Institute Bioscience, Norfolk, UK

[17]Earlham Institute, Norwich Research Park, Norfolk, UK

[18]Institut de Biologie Intégrative et des Systèmes, Université Laval, Québec, Québec, Canada

[19]Center for Forest Mycology Research, Northern Research Station, US Forest Service, Madison, Wisconsin, USA

[20]GenPhySE, Université de Toulouse, INRAE, ENVT, Castanet Tolosan, France

[21]Quantitative Biology Center (QBiC), University of Tübingen, Tübingen, Germany

[22]Laboratory of Environmental Microbiology, Institute of Microbiology of the Czech Academy of Sciences, Praha, Czech Republic

[23]KU Leuven, Department of Microbiology, Immunology and Transplantation, Rega Institute for Medical Research, Laboratory of Molecular Bacteriology, Leuven, Belgium

Correspondence
Sten Anslan, Institute of Ecology and Earth Sciences, University of Tartu, Tartu, Estonia.
Email: sten.anslan@ut.ee

Present address
Jonathan M. Palmer, Genencor Technology Center, IFF, Palo Alto, California, USA

Funding information
Deutsche Forschungsgemeinschaft (DFG), Grant/Award Number: EXC2124; European Regional Development Fund and the programme Mobilitas Pluss, Grant/Award Number: MOBTP198; Genome Canada and Ontario Genomics; Grantová Agentura České Republiky, Grant/Award Number: 21-17749S; National Institute of General Medical Sciences, Grant/Award Number: P20GM103449

Handling Editor: Simon Creer

Abstract

Environmental DNA (eDNA) metabarcoding has gained growing attention as a strategy for monitoring biodiversity in ecology. However, taxa identifications produced through metabarcoding require sophisticated processing of high-throughput sequencing data from taxonomically informative DNA barcodes. Various sets of universal and taxon-specific primers have been developed, extending the usability of metabarcoding across archaea, bacteria and eukaryotes. Accordingly, a multitude of metabarcoding data analysis tools and pipelines have also been developed. Often, several developed workflows are designed to process the same amplicon sequencing data, making it somewhat puzzling to choose one among the plethora of existing pipelines. However, each pipeline has its own specific philosophy, strengths and limitations, which should be considered depending on the aims of any specific study, as well as the bioinformatics expertise of the user. In this review, we outline the input data requirements, supported operating systems and particular attributes of thirty-two amplicon processing pipelines with the goal of helping users to select a pipeline for their metabarcoding projects.

KEYWORDS
amplicon data analysis, bioinformatics, environmental DNA, metabarcoding, pipeline, review

1 | INTRODUCTION

Advances in high-throughput sequencing (HTS) technologies have boosted the application of molecular methods for species identifications. Metabarcoding, the simultaneous tagging, sequencing, and identification of multiple species within a single environmental sample (Taberlet, Coissac, Hajibabaei, & Rieseberg, 2012) is now a widely applied technique in biodiversity research (Compson et al., 2020). Metabarcoding involves PCR-based amplification of taxonomically informative gene fragments ('DNA barcodes', markers) that are subsequently sequenced to be used for species identifications in the presence of reference sequence data (DNA barcodes). Before identification, the sequencing data are processed in several steps (Figure 1) where one of the first steps is usually performing quality control on the data. A sequence analysis pipeline is generated by applying various steps using a collection of software and algorithms with the ultimate goal of producing an accurate features table with potential taxon annotations by sample (i.e. with features metadata). In metabarcoding, features refer to amplicon sequence variants (ASVs), operational taxonomic units (OTUs) or annotated taxa; and their sample-wise distribution matrix can be further utilized in relevant biostatistical analyses.

With the emergence of practical guidelines (e.g. Bruce et al., 2021; Lear et al., 2018; Tedersoo et al., 2022), the scalability and throughput of environmental DNA (eDNA; a mixture of DNA from different organisms in an environmental sample; Taberlet, Coissac, Pompanon, et al., 2012) sample processing has contributed to the popularity of the metabarcoding approach among ecologists. However, one of the bottlenecks of metabarcoding is choosing how to process sequencing data sets into relevant feature tables bioinformatically. Among the first highly successful software developed for that purpose have been mothur (Schloss et al., 2009), USEARCH (Edgar, 2010) and QIIME 1 (Caporaso et al., 2010), which consists of algorithms that can be combined to create full metabarcoding data analysis pipelines. Over the years, these programs have been supplemented with additional algorithms to help reduce artefactual sequences and implement different sequencing clustering approaches. These pipelines were initially developed for microbial 16S rRNA amplicon analysis, but the applications of metabarcoding have been expanded to a wide range of taxa from various environmental samples, resulting in a boom in pipeline development. Some workflows include a set of newly designed algorithms, but others represent a combination of different open-source tools used for the different analysis steps bound into executable pipelines. From the lack of easy-to-use bioinformatics tools from the early age of metabarcoding, we have reached a phase where the choices are so numerous that it may be difficult to select among the multitude of analytical workflows.

Below, we delve into the properties of thirty-two software packages that can be used for the bioinformatics processing of metabarcoding data. In this review, we outline several key aspects of those metabarcoding software, including which ones represent software suites or precompiled pipelines, consideration of the software depending on the utilized sequencing platform, available operating system and interface preference (Figure 2). By addressing these components, we seek to offer a comprehensive understanding of the software landscape for metabarcoding projects. Since different users will have different needs, we do not seek to recommend

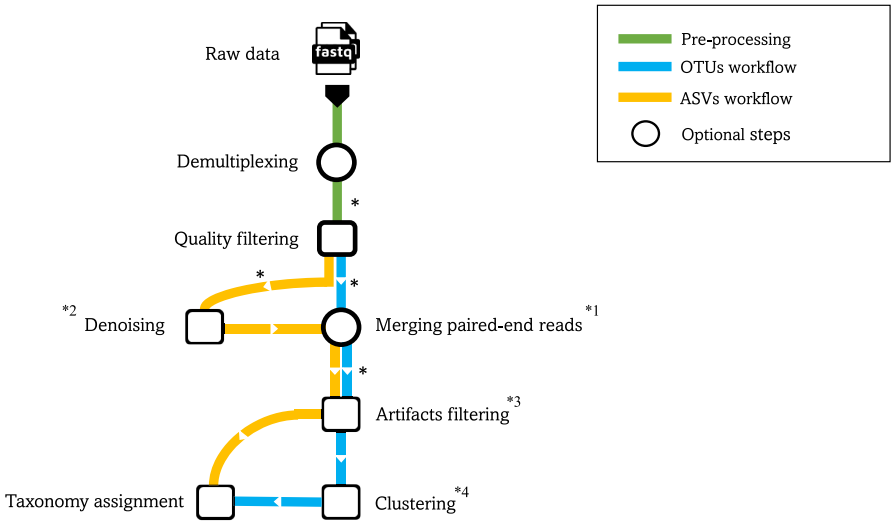


FIGURE 1 Examples of basic bioinformatics workflows for metabarcoding data. The workflow begins with demultiplexing, assigning reads to respective samples based on unique molecular identifiers. Next, quality filtering removes low-quality reads to reduce errors and improve reliability. Denoising algorithms identify and correct sequencing errors while preserving biological variation. For paired-end reads, merging combines forward and reverse reads into single-end sequences. Artifacts filtering aims to remove artifacts such as chimeras and NUMTs. Clustering groups of sequences into features. Finally, taxonomic assignment of the features against a reference database. * Primer trimming between any of these steps can be applied. *1 Only for paired-end data (may be performed before or after quality filtering). *2 Error correction; formation of ASVs. *3 Including chimera filtering, off-target gene removal (pseudogene removal, ITS extraction). *4 Formation of OTUs/swarm-clusters.

the best-performing pipeline, a task that would be highly context-dependent, but rather to give an overview of the software that are available for metabarcoding data analysis. Table 1 lists the software discussed here and their extended description and specific capabilities are outlined in Appendix S1.

2 | SOFTWARE SUITES AND PRECOMPILED PIPELINES

The metabarcoding data processing software may be roughly divided into two categories based on their structure for executable algorithms—software providing a set of algorithms, and pipelines providing a predefined chain of algorithms. USEARCH, VSEARCH (Rognes et al., 2016), DADA2 (Callahan et al., 2016), OBITools (Boyer et al., 2016), mothur and QIIME 2 (Bolyen et al., 2019) are software suites that host numerous algorithms for sequence data analysis, thus are highly customizable to construct user-defined pipelines with a specific chain of commands and settings. VSEARCH largely mirrors the diverse functionalities of USEARCH, but without the requirement to purchase a licence for a version that can handle large data sets and use more than 4 GB of a computer's memory. Besides consisting of a large set of unique data processing algorithms, mothur and QIIME 2 wrap some functionalities of VSEARCH and/or DADA2.

Software providing a predefined chain of algorithms represents full analytical pipelines with specific workflow steps, as depicted in Figure 1. The predefined pipelines consist of workflow steps validated on certain sequencing data to facilitate the metabarcoding data analysis, which may be especially convenient for users with few bioinformatics skills. Some workflows include a set of newly designed algorithms, but others represent a combination of different open-source tools used for the different steps that are bound into easily executable pipelines Although, these pipelines are predefined, they often allow the user to customize the settings depending on the characteristics of the sequencing data set.

3 | BASIC STRUCTURE OF A METABARCODING PIPELINE

3.1 | Demultiplexing

Demultiplexed sequences are often provided to users since this process has been integrated into sequencing provider software, such as bcl2fastq (for Illumina raw data) and SMRT Tools (for Pacific Biosciences [PacBio] reads). Demultiplexing distributes the sequences into individual files, most often corresponding to the experiment's samples. However, when requesting multiplexed data (pooled sequences from multiple samples), the reads from a sequencing run may need to be demultiplexed before using some of the application software. The demultiplexing step is not incorporated into all software (Table 1). In cases where it is not, other programs such as cutadapt (Martin, 2011), sdm (Falk Hildebrand et al., 2014) or lima (https://lima.how/; for single-end reads only) may be used. In cases where multiple markers are used per sample (via multiplex PCR), amplicons from different primer sets should also be split. The latter step is included in the Anacapa and VTAM pipeline

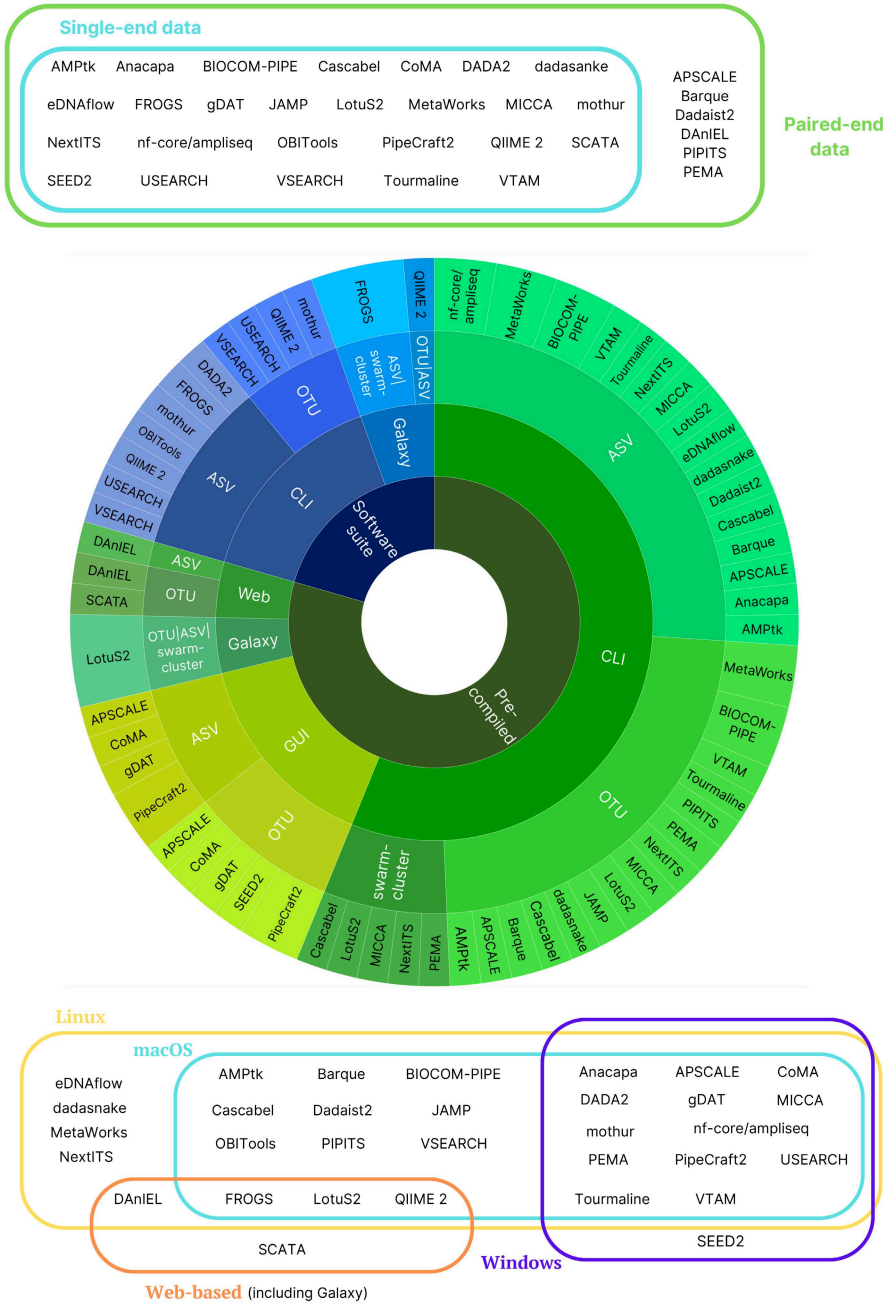


FIGURE 2 Software for metabarcoding data bioinformatics processing categorized by input read type (paired-end, single-end (the tools in electric blue are capable of handling both paired-end and single-end reads)), software type (suite, precompiled pipeline), interface (CLI, GUI, Web, Galaxy web platform), produced feature type (OTU, ASV, swarm-cluster) and operating system (Linux, macOS, Windows).

(Curd et al., 2019; González et al., 2023), where different markers are automatically separated based on the primer sequences.

3.2 | Primer trimming

Sequencing adapters, indexes and primers should be removed before the following analyses. Depending on the data structure, the former two may be absent, but the programs mentioned above (for demultiplexing) may be used to double-check this and remove primers. Adapter/primer clipping is often implemented into a pipeline by wrapping the cutadapt, Trimmomatic (Bolger et al., 2014) or AdapterRemoval (Lindgreen, 2012) functionality, in others done during quality filtering and demultiplexing steps (e.g. sdm; Table S1).

3.3 | Quality filtering & merging paired-end reads

The following phases of a standard DNA metabarcoding pipeline is sequence filtration based on the read quality scores, removal of putative chimeric/artefactual sequences, the definition of features (e.g. ASVs, OTUs) and taxonomic annotation of the features (Figure 1). In the case of paired-end data, the merging process of the overlapping sequences may be performed before or after the quality filtering step and even sometimes after the sequence clustering step. There are a multitude of strategies for performing the above-listed processes, where the selection of an approach may depend on the specific characteristics of the sequencing data or the aims of the study. The strategies for quality filtering include per-sequence or per-nucleotide(s)-based filtering. Per-sequence filtering includes

TABLE 1 A list of the reviewed metabarcoding data analyses pipelines (in alphabetical order).

Pipelines	First/latest release	Software type	Feature	Demux	Primer/adapter removal	Marker
AMPtk	Nov 2017/Jan 2023	Precompiled	ASV, OTU	Yes	Yes	16S, 28S, COI, ITS
Anacapa	June 2018/June 2018	Precompiled	ASV	Multilocus demux based on primers	Yes	Multimarker
APSCALE	Jan 2022/Feb 2023	Precompiled	ASV, OTU	No	Yes	Multimarker
Barque	Sep 2016/Dec 2022	Precompiled	ASV, OTU	No	Yes	Multimarker
BIOCOM-PIPE	Aug 2020/Jan 2021	Precompiled	ASV, OTU	Yes	Yes	16S, 18S, 23S
Cascabel	Nov 2020/Dec 2022	Precompiled	ASV, OTU, swarm-cluster	Yes	Yes	Multimarker
CoMA	Dec 2020/Jan 2022	Precompiled	ASV, OTU, swarm-cluster	No	Yes	Multimarker
DADA2	May 2016/Nov 2022	Set of algorithms	ASV	No	Yes	Multimarker
Dadaist2	May 2020/Jul 2022	Precompiled	ASV	No	Yes	Multimarker
dadasnake	Apr 2020/Feb 2023	Precompiled	ASV, OTU	No	Yes	Multimarker (with ITSx)
DAnIEL	Apr 2021/May 2021	Precompiled	ASV	Yes	Yes	ITS
eDNAflow	Jul 2020/Jan 2021	Precompiled	ASV	Yes	Yes	Multimarker
FROGS	Jan 2016/Apr 2023	Set of algorithms	ASV, swarm-cluster	Yes	Yes	Multimarker (with ITSx)
gDAT	Mar 2020/Jan 2021	Precompiled and set of algorithms	OTU	No	Yes	18S, ITS
JAMP	Jan 2017/Jan 2022	Precompiled	OTU	Yes	Yes	Multimarker (without taxonomic assignment)
LotuS2	May 2021/Apr 2023	Precompiled	ASV, OTU, swarm-cluster	Yes	Yes	Multimarker (16S, 18S, 23S, 28S, ITS with ITSx)
MetaWorks	Jun 2020/Mar 2023	Precompiled	ASV, OTU	No	Yes	Multimarker; SSU (12S, 16S, 18S), ITS (with ITSx), LSU (28S), COI (with pseudogene removal), rbcL (with pseudogene removal)
MICCA	Oct 2014/Aug 2018	Precompiled	ASV, OTU, swarm-cluster	Yes	Yes	16S, 18S, 28S, ITS
mothur	Feb 2009/May 2022	Set of algorithms	ASV, OTU	Yes	No	Multimarker
NextITS	July 2022/May 2023	Precompiled	ASV, OTU, swarm-cluster	Yes	Yes	ITS (with ITSx)
nf-core/ampliseq	Dec 2018/Mar 2022	Precompiled	ASV	No	Yes	Multimarker (with ITSx)
OBITools3	Sep 2019/Sep 2022	Set of algorithms	ASV	Yes	Yes	Multimarker
PEMA	Feb 2020/Dec 2021	Precompiled	OTU, swarm-cluster	No	Yes	Multimarker

(Continues)

TABLE 1 (Continued)

Pipelines	First/latest release	Software type	Feature	Demux	Primer/adapter removal	Marker
PipeCraft2	Dec 2021/Dec 2022	Precompiled and set of algorithms	ASV, OTU	Yes	Yes	Multimarker (with ITSx and pseudogene removal)
PIPITS	Dec 2014/Nov 2022	Precompiled	OTU	No	Yes	ITS (with ITSx)
QIIME 2	July 2016/May 2023	Set of algorithms	ASV, OTU	Yes	Yes	Multimarker (with ITSxpress)
SCATA	Feb 2010/Oct 2021	Precompiled	OTU	Yes	Yes	Multimarker (with ITSx)
SEED2	Feb 2018/Oct 2020	Precompiled	OTU	Yes	Yes	16S, ITS (with ITSx)
Tourmaline	July 2022/Apr 2023	Precompiled	ASV, OTU	Yes	Yes	Multimarker
USEARCH	2010/Mar 2020	Set of algorithms	ASV, OTU	No	No	Multimarker
VSEARCH	Nov 2014/Sep 2022	Set of algorithms	ASV, OTU	No	No	Multimarker
VTAM	Oct 2020/May 2022	Precompiled	ASV, OTU	Yes	Yes	Multimarker (with pseudogene removal)

Note: Column "software type," denotes the structure of the pipeline. "Feature" denotes the pipeline output unit, operational taxonomic units (OTU), amplicon sequence variants (ASV) or swam-clusters (using swarm). "Demux" indicates whether the demultiplexing step is implemented. "Primer/adapter removal" indicates whether primer/adapter removal step in implemented. Column "Marker" states the gene region for which the pipeline has been benchmarked. More details about software in Appendix S1.

discarding the whole sequence if it does not meet the threshold requirements, whereas the per-nucleotide(s) approach truncates the sequence from the position below the threshold to keep a partial amplicon. Among the quality threshold calculation methods, the filtering based on the expected number of errors (sum of the error probabilities) is preferred over the average quality score threshold (Edgar & Flyvbjerg, 2015), because a 'good' average quality score may mask several bases with relatively high error probabilities that can subsequently propagate into false positive features. Haplotype-level (ASV) analyses may require relatively stringent quality cutoffs for accurate fine-resolution analyses, whereas cutoffs may be more lenient when generating OTUs (because clustering collapses many of the accumulated errors during sequencing) or if summarizing data to more inclusive taxonomic ranks (species, genera, etc.).

3.4 | Artefacts filtering

Putative chimeric sequences are most commonly removed by comparing sequences against each other (de novo method), but with the existence of an appropriate (curated, chimera-free) reference database, additional reference-based chimera filtering is recommended (Tedersoo et al., 2022). De novo methods tend also to discard sequences that are incorrectly flagged as chimeric (false-positive chimeric sequences; Pauvert et al., 2019; Tedersoo et al., 2022). The loss of these false positive chimeric sequences detection may be more 'costly' for data sets with low sequencing depths. To attempt to rescue those real members of the sequenced community (false-positive chimeras), NextITS (Mikryukov et al., 2022) and FROGS (Bernard et al., 2021; Escudié et al., 2018) pipelines have implemented an approach to recover sequences that occur in multiple samples (because the formation of an identical chimera in different PCR runs is highly unlikely). With NextITS, it is also possible to inspect the distribution of UCHIME scores (Edgar et al., 2011) of putative chimeras, which allows adjustment of sensitivity-specificity trade-offs in chimera discrimination according to the study aims. A custom false-positive chimeras recovery method is also implemented in BIOCOM-PIPE (Djemiel, Dequiedt, et al., 2020), where initially discarded chimeras can be recovered based on their taxonomic assignments.

3.5 | Denoising & clustering

The formation of features in many pipelines includes both ASVs and OTUs (Table 1). ASVs are identical denoised reads with as few as 1 base pair difference between variants, representing an inference of the biological sequences prior to amplification and sequencing errors (Callahan et al., 2017). ASVs are mainly formed through the two most popular denoising algorithms, DADA2 and UNOISE (Edgar, 2016b). Although features formed via UNOISE are referred as zOTUs (zero-radius OTUs; Edgar, 2016b), sometimes also as ESVs (exact sequence variants; Buchner et al., 2022; Porter & Hajibabaei, 2022), we herein denote those with a unified term—ASVs. Although less frequently implemented in the pipeline, deblur (Amir et al., 2017) and obiclean (Boyer et al., 2016) are other denoising algorithms for ASVs formation (Table S1). The OTU clustering approaches include a much wider set of algorithms across different software (Table S1), which typically rely on global sequence similarities. Notably, the clustering process in SCATA (Durling et al., 2011) includes collapsing of homopolymer

regions to account for homopolymer-length errors during sequencing (which are especially common on 454 and Ion Torrent platforms; Laehnemann et al., 2016). Similarly, before the formation of features, NextITS implemented the correction of homopolymer errors in PacBio reads. Swarm (Mahé et al., 2022) is a notably different sequence clustering approach. It relies on the maximum number of differences between reads (local linking threshold), where clusters are resilient to input-order changes, therefore, forming stable, high-resolution features (herein referred to as swarm-clusters). Swarm is currently implemented in Cascabel, CoMA, FROGS, LotuS2 (Özkurt et al., 2022), MICCA (Albanese et al., 2015), NextITS, and PEMA (Zafeiropoulos et al., 2020) pipelines (Table S1; Appendix S1).

Which type of features to prefer may be context-dependent, and both may even be used in the same study. Denoised ASVs provide a biologically informative fine-scale resolution that are collapsed during the OTU formation process (Callahan et al., 2016). For example, by testing several ASVs and OTUs-based workflows for detecting the *Botrylloides* (Ascidiacea) haplotypes, Couton et al. (2021) reported that ASVs pipeline (DADA2) retrieved all expected haplotypes, whereas OTUs datasets (99.5% threshold for clustering) missed several expected haplotypes by collapsing very closely related ones into a single OTU. By default, denoisers tend to discard low-abundant sequence variants, which are more likely to be artefacts (Anslan et al., 2021; Reitmeier et al., 2021). Although denoising greatly lowers the fraction of spurious features (e.g. de Santiago et al., 2022), in some contexts it may be difficult to separate noise from a real signal in low abundant ASVs. For example, the denoising process might discard some rare taxa, that is, ASVs with a low number of sequences (Edgar, 2016b; Nearing et al., 2018). This may have a larger impact when working with a data set with a relatively low sequencing depth. Nevertheless, in some pipelines (e.g., DADA2, FROGS, VSEARCH, and USEARCH), the sensitivity to rare ASVs can be modified according to the user's needs. Importantly, ASVs represent stable and reproducible units across studies whereas OTUs are dataset-specific features (Callahan et al., 2017). However, the ASVs approach may not accurately reflect species composition in the community of, for example, metazoans with highly variable levels of intraspecific polymorphism in the COI gene (Brandt et al., 2021) and fungi with multiple different ITS copies per genome and their size polymorphism (Estensmo et al. 2021; Tedersoo et al., 2022) except when the treatment of ITSs is particularly taken into account (as, e.g., in FROGS; Bernard et al., 2021). If relevant, upon formation of ASVs, those may be subjected to further clustering (Antich et al., 2021; Brandt et al., 2021; Porter & Hajibabaei, 2020). The latter approach is implemented in, for example, MetaWorks (Porter & Hajibabaei, 2022), PipeCraft2 (Anslan et al., 2017), and dadasnake (Weißbecker et al., 2020). Additionally, QIIME 2, nf-core/ampliseq (Ewels et al., 2020; Straub et al., 2020) and LotuS2 support the features collapsing by annotated taxon levels, resulting in taxa features. Overall, the resulting community patterns of a study are often highly similar regardless of the utilized feature (e.g. Glassman & Martiny, 2018; Kang et al., 2021; Porter & Hajibabaei, 2020), but may vary in recovering rare taxa (Nearing et al., 2018).

After the formation of features, the presence of a long tail of low-abundant units is common. This tail is often discarded, assuming that a large proportion of low-abundant features are artefactual (Huse et al., 2010; Reeder & Knight, 2009). However, without applying arbitrary cutoff levels (e.g. removing features with <10 reads per-sample; Brown et al., 2015), the postclustering process aids in removing the erroneous features but keeping the rare, potentially real ones. Postclustering tools, such as LULU (Frøslev et al., 2017) are implemented in AMPTk (Palmer et al., 2018), eDNAflow (Mousavi-Derazmahalleh et al., 2021), APSCALE (Buchner et al., 2022), LotuS2, PipeCraft2 and ReClustOR (Terrat et al., 2020) in BIOCOM-PIPE.

Postclustering, however, does not resolve the tag-switching phenomena, where some low-abundant nonartificial features may represent false-positive occurrences across samples. Tag-switching is a well-documented issue (e.g. Carlsen et al., 2012; Rodriguez-Martinez et al., 2022; Schnell et al., 2015), but is rarely considered in practice because the low proportions of tag-switching errors do not heavily impact the community-level analyses (e.g., Anslan et al., 2021). Nevertheless, the incorrect sample assignments of features artificially inflate the richness. For discarding potential tag-switching errors from the feature table, pipelines such as NextITS, LotuS2, PipeCraft2 and Dadaist2 (Ansorge et al., 2021) wrap the UNCROSS2 (Edgar, 2018b) algorithm (from USEARCH). Based on the included control samples, AMPtk and VTAM attempt to automatically correct for tag-switching errors. Notably, the tag-switching issue can be minimized by accounting for this in the laboratory work protocol (Carøe & Bohmann, 2020; Taberlet et al., 2018). However, for further feature occurrence filtering to filter out low-confidence detections biological/technical replicates per sample are recommended (Gold et al., 2021). This allows examining the feature co-occurrence patterns across replicates to estimate detection probabilities and retain only high-confidence detections (by applying, e.g., site occupancy modelling). Among the precompiled pipelines, VTAM implements a feature occurrence filtering procedure based on the user-defined number of technical replicates they appear in, and samples may be discarded when the sequence composition in the replicate samples is too dissimilar. Not incorporated to the pipelines discussed here, but the MetabaR package (Zinger et al., 2021) aids to detect different types of artefactual sequences, such as potential contaminants, tag-switches, and dysfunctional PCRs (on the basis of similarities between replicate samples).

3.6 | Taxonomy assignment

In the reviewed pipelines, the most common taxonomy assignment methods include alignment-based (such as BLAST; Altschul et al., 1997) and sequence composition-based approaches (e.g. RDP Naïve Bayesian classifier; Wang et al., 2007; see Table S1). Several studies have tested the accuracy of different taxonomy assignment methods (e.g. Bokulich et al., 2018; Curd et al., 2019; Edgar, 2018a; Hleap et al., 2021; Richardson et al., 2017) and have recognized a relationship between the reference database completeness and

the classification accuracy. Regardless of the taxonomic group, the reference databases are far from being complete (Gold et al., 2021; McGee et al., 2019; Nilsson et al., 2016; Weigand et al., 2019). Therefore, a trade-off between the detection of true-positives (correctly assigned sequences) and false-positives (incorrectly assigned sequences), that is, the precision and the recall rate, should be considered when choosing a threshold for the classification (Bokulich et al., 2018; Edgar, 2018a). Hleap et al. (2021) suggested that a multilayer approach could enhance the effectiveness of similarity-based methodologies. The goal of this strategy is to improve the precision of taxonomic assignments, minimize the occurrence of false positives, and boost the efficiency of the classification process. VTAM's taxonomy assignment function has incorporated elements of this strategy. It begins the assignment process with a high percentage identity threshold, which is gradually lowered until the lowest taxonomic group is established.

Although composition-based (and other 'complex') methods may be more sensitive to the patchy coverage databases than 'simple' alignment-based methods (Hleap et al., 2021), in certain circumstances Naïve Bayesian classifiers may outperform BLAST (Rosen et al., 2011). However, the assignment accuracy to higher taxonomic ranks (such as Family level) generally has similar performance across the approaches (Hleap et al., 2021). A recent development in QIIME 2 involves utilizing public microbiome data for probabilistic taxonomy assignment (Kaehler et al., 2019). This method offers several advantages, including the potential for higher resolution taxonomic classification for instance, it can enable species-level classification when previously only genus-level classification was possible. Other pipelines, such as LotuS2, can assign features from multiple taxonomic databases, to preferentially assign taxonomies based on databases that are specific to a given environment. FROGS returns an original multiaffiliation output to highlight databases conflicts and uncertainties taxonomic affiliations. AMPtk implements a hybrid taxonomy assignment that utilizes global alignment (VSEARCH) and SINTAX (Edgar, 2016a) to calculate a consensus LCA (last common ancestor) taxonomy. Regardless of the taxonomy assignment methods used, the reference database should also include a proportion of nontarget taxa (including potential contaminants) to limit the overclassification of features to the target taxa (e.g. Anslan et al., 2018).

4 | SEQUENCING PLATFORM

The most commonly utilized high-throughput sequencing approaches for metabarcoding are short-read, second-generation technologies, such as those provided by Illumina platforms. These platforms produce a high number of high-quality paired-end short reads (up to 300 bp for single-end) with a relatively low cost per sample. Therefore, most amplicon data analysis pipelines are set up to be able to handle paired-end sequencing data (Table 1; Figure 2). As the MGI-Tech platforms may also produce paired-end reads (with comparable data quality and throughput properties compared with Illumina; Anslan et al., 2021), the paired-end compatible pipelines may be used to analyse data from the latter platforms as well.

Some pipelines are restricted to paired-end input, that is, the analytical pipeline cannot be completed using only a single-end part of the data, or sequencing data from the long-read (third-generation) sequencing platforms (Table 1; Appendix S1). With the rapid developments in third-generation sequencing accuracy and throughput, there is increasing interest to generate longer metabarcodes, which potentially increases the taxonomic resolution (Tedersoo et al., 2021, 2022) and has lower sequencing bias toward short amplicons (Castaño et al., 2020). Therefore, some software developed for short reads have been updated to also process longer sequences (specifically PacBio reads; Appendix S1). Although, some of the software considered here have performed well for sequencing data (HiFi reads) processing from PacBio platforms (e.g., Castaño et al., 2020; Heeger et al., 2018; Tedersoo & Anslan, 2019), the data from Oxford Nanopore Technologies (ONT) platform may require other customized approaches (Baloğlu et al., 2021). Herein listed software (Table 1) have not been specifically developed for analysing ONT data, thus care should be taken when applying these tools for nanopore reads.

The sequencing depth may vary considerably between sequencing platforms. For example, Illumina MiSeq system may produce up to 25 M 2×300 bp reads and NovaSeq up to 1600 M 2×250 bp reads per flow cell, whereas the throughput of PacBio Sequel II(e) system is up to 4 M HiFi reads. Since the denoising tools are sensitive (by default) to low abundant sequences, then one must be wary that strict denoising of low sequencing depth samples increases the number of false negatives, that is, rare true positives may be denoised out (Furneaux et al., 2021), especially if the samples contain complex communities, such as found in soil. Besides sequencing depth, the detection of rare sequence variants may be affected by the different chemistry utilized by different platforms (e.g. NovaSeq vs. MiSeq; Singer et al., 2019). Importantly, denoising algorithms, such as UNOISE and deblur, are designed for Illumina reads and may not perform well with data from other sequencing platforms. Therefore, opting for an OTU clustering approach may be more appropriate for analysing complex communities sequenced by the third-generation platforms. The remaining bioinformatically unscreened low-abundance spurious OTUs could then be abandoned after the post-clustering step, by, for example, filtering out unclassified features at the phylum level, and based on the number of samples they occur in (e.g. discard features only observed in one sample). Although, DADA2 has a specific denoising function to estimate errors from PacBio reads (Callahan et al., 2019) which performs well also on synthetic long reads (Callahan et al., 2021), its application may still require higher sequencing depth for high diversity samples (Furneaux et al., 2021). However, the throughput of the most recent PacBio long-read sequencing system, Revio (commercially available from the first half of 2023), is expectedly up to 15 times higher compared with Sequel II. But the performance of the denoisers with the greatly increased throughput of long-read data (exceeds the throughput of, e.g., Illumina MiSeq) is yet to be tested.

In case the amplicon is shorter than the sequencing cycle (e.g. expected amplicon is ~130 bp, but one cycle synthesizes 250 bp), the Illumina NovaSeq and NextSeq platforms may extend the amplicon by adding a poly-G tail (with 'good' quality scores). Therefore, trimming primers from amplicon reads should be used by default, as this will discard the overhanging sequence parts. Fastp tool (Chen et al., 2018), wrapped also in PipeCraft2 and NextITS, may be used to specifically trim these non-biological poly-G (or poly-X) tails. Additionally, Phred scores from third-generation sequencing platforms range from 0 to 93, thus may require adjustments of the maximum quality score setting when using, for example, VSEARCH or USEARCH software (where the default is 41, for Illumina).

5 | OPERATING SYSTEMS AND WORKFLOW MANAGERS

Unix-based operating systems (OS), such as Linux and macOS, are the most common and convenient platforms in bioinformatics, as users may run software with a comparable interface on a personal computer or high-performance computing (HPC) system. As a result, they are by far the most widely used for the development and use of bioinformatics tools. Accordingly, almost all the presented pipelines can be executed in Linux and/or macOS operating systems (Table 1; Figure 2). Since many users have computers running on Windows-based operating systems, several pipeline developers have gone through the effort of making the Unix-based workflows executable in Windows; sometimes through native code adaptations or by making their software available in either containers or through websites (such as Galaxy; Galaxy Community, 2022), making the pipelines independent of the OS (Table 1).

Some metabarcoding data analysis tools, such as DADA2 rely exclusively on R and are thus also compatible with any OS that can execute R. JAMP (https://github.com/VascoElbrecht/JAMP) is another R package that wraps full metabarcoding and haplotyping pipelines, although it is only available for Linux and macOS. Additionally, with the development of containerization technology (e.g. Docker, Singularity), it becomes easier to develop bioinformatics pipelines that can run on the three major operating systems, Windows, macOS, and Linux. A container encapsulates the code and dependencies needed for the data analyses so that the pipeline may run reliably on any OS. Once the containerization software is installed, users are free to instal all the underlying dependencies. For a few of the presented pipelines, the developers have included the prebuilt containers and/or virtual machine images required to run it (Table 1; Appendix S1). Pipelines such as those distributed by nf-core/ampliseq, PipeCraft2, PEMA, and Tourmaline require utilizing Docker/Singularity containers at the backend, so the core bioinformatics processes are running on a Linux environment but may also be executed on Windows and macOS systems. Moreover, containerized pipelines resolve the numerical instability issue occurring while running software on different computational platforms (Di Tommaso et al., 2017), ensuring the consistency of results and allowing more reproducible computational workflows.

Essentially all the pipelines can be run on any OS via containers or virtual machines. However, containers are preferred to virtual machines (e.g. VirtualBox), as virtualization (i.e. running a second OS on top of the main OS) has high overhead and comes at the cost of a computer's RAM usage, which ultimately limits the amount of data that can be processed. Considering container engines, Docker is usually unavailable on HPC clusters, as potential vulnerability could provide means to gain root access to the system they are running on. Therefore, Singularity (Kurtzer et al., 2017) is generally more widespread on HPC clusters as it was specifically developed for it.

With the capacity to provide computational resources, web-based platforms, such as DAnIEL (Loos et al., 2021) and SCATA may be simply used through a web browser on any operating system. Additionally, some other pipelines, such as FROGS and LotuS2, can also be accessed through Galaxy websites, and nf-core/ampliseq (Straub et al., 2020) can be launched via Nextflow Tower (a monitoring and management platform for Nextflow workflows).

The increasing complexity of bioinformatics pipelines, which consist of a large number of computational steps, encouraged the development of workflow management systems capable of orchestrating in a scalable and reproducible manner (Mölder et al., 2021; Wratten et al., 2021). Workflow managers allow pipelines to resume after a failure and start from the last successfully completed step, automate pipeline execution triggered by input or reference data updates and perform parameter exploration. Nextflow (Di Tommaso et al., 2017) and Snakemake (Koster & Rahmann, 2012; Mölder et al., 2021) are among the most prominent workflow management systems in the field of bioinformatics. They simplify pipeline development, maximize resource usage efficiency and handle installation and versioning of the software dependencies (e.g. using Docker and Singularity containers or conda environments). These systems allow running workflow steps in parallel locally or using resources of HPC clusters or commercial cloud computing providers (Amazon web services (Bai et al., 2019), Microsoft Azure (Copeland et al., 2015), Google Cloud (Hussain & Aleem, 2018)) almost without the need to adapt pipeline code to a specific platform architecture. MetaWorks, dadasnake, Tourmaline (Thompson et al., 2022), and Cascabel (Asbun et al., 2020) are examples of Snakemake-based pipelines, while nf-core/ampliseq, eDNAflow and NextITS were developed using Nextflow.

6 | THE INTERFACE

Generally, the Unix-based command-line interfaces (CLI; commands are typed into a terminal) are often preferred by analysts with bioinformatics experience. That is because most of the pipelines are developed as CLI-runnable software that can be operated on HPC clusters, but also due to the flexibility and availability of applying various custom processes to manage the data effectively. Although the CLI tools offer numerous advantages, using a CLI might be intimidating for users with less programming experience. To facilitate the analysis of metabarcoding data by nonbioinformaticians, APSCALE,

CoMA (Hupfauf et al., 2020), gDAT (Vasar et al., 2021), PipeCraft2 and SEED 2 (Vetrovský et al., 2018) provide a graphical user interface (GUI; interaction via clickable graphical icons; Table 1; Figure 2) as a front-end for specifying the settings of the bioinformatics analyses, which will be executed on the back-end. Depending on the architecture, the GUI-based applications may require more RAM than CLI pipelines. Pipelines that have web server support (DAnIEL, SCATA) or have been implemented into Galaxy server (LotuS2, FROGS, QIIME 2) naturally possess a web-based GUI for specifying the settings of the analysis. Some software that is wrapped into GUI may also be executed through CLI (Table 1).

7 | MARKER-SPECIFIC PIPELINES

A marker (i.e., 'DNA barcode') is a taxonomically informative gene fragment that is utilized for species identifications in the presence of reference sequence data. Bioinformatics processes combined in a pipeline may be specifically designed to analyse amplicons from a specific marker, that is, the analytical steps may depend on the characteristics of the amplicons. For example, when processing ITS amplicon data, it is common to remove conservative flanking genes of ITS for accurate taxonomic classification purposes (Tedersoo et al., 2022; Vu et al., 2022). When processing sequences from the COI gene, removing the co-amplified putative nuclear mitochondrial pseudogenes (NUMTs) is highly recommended (Creedy et al., 2022; Porter & Hajibabaei, 2021; Song et al., 2008). The subsections below outline the herein-considered marker-specific and multimarker pipelines and highlight some of the results from their benchmarking trials.

7.1 | Prokaryotic 16S rRNA

Amplicon sequencing targeting the 16S rRNA gene is commonly utilized to investigate microbiomes from various ecosystems/substrates (Knight et al., 2018; Pollock et al., 2018; Staats et al., 2016). The 16S gene sequence is roughly 1500 bp in length and contains nine distinct hypervariable regions (V1–V9). The V4 hypervariable region is most often used in short-read sequencing, whereas full-length 16S analyses are becoming increasingly utilized with the increased quality, availability, and decreasing costs of long-read sequencing methods. For processing 16S amplicons, mothur, USEARCH, QIIME 2 and DADA2 are the most used ones. Recently established pipelines such as dadaist2, dadasnake, nf-core/ampliseq, Tourmaline also wrap QIIME 2 and/or DADA2 functionalities and are thus optimized for 16S (but not exclusively) analyses. BIOCOM-PIPE, Cascabel, CoMA, LotuS2, MICCA, PEMA and FROGS have also benchmarked their pipelines using 16S data sets. However, since the bioinformatics processing of 16S amplicon data was at the forefront of metabarcoding data analyses before the wide-scale utilization of other markers, other multimarker pipelines (Table 1) that consist of critical filtering steps may also be used to process 16S reads.

Testing different workflows on 16S V4 amplicon mock data (known composition of taxa in a sample), Straub et al. (2020) found QIIME 2 pipeline with the DADA2 plugin being the most optimal compared to mothur, QIIME 1 and MEGAN (Huson et al., 2007) workflows. Based on the benchmarking results, the nf-core/ampliseq pipeline was developed which demonstrated a high degree of similarity with the results produced by QIIME 2. Prodan et al. (2020) reported good performance of all tested ASV workflows (DADA2, QIIME 2 deblur, UNOISE3), but with slight variations in their sensitivity and specificity to detect mock community members. In the latter study, two OTU workflows also performed well (UPARSE, mothur; but not QIIME 1-uclust), but with lower specificity than ASV pipelines. A more recent study by Özkurt et al. (2022) reported the higher accuracy of LotuS2 compared with QIIME 2, DADA2, and PipeCraft2. The LotuS2 pipeline runs with stringent read filtering and implements a unique feature, a 'seed extension' algorithm, that improves the quality of a feature's representative sequence. By introducing the CoMA pipeline that uses LotuS1/2 (Hildebrand et al., 2014) at its core, Hupfauf et al. (2020) reported a good performance of all tested pipelines (CoMA, QIIME 2, mothur). However, some degree of variability was evidently depending on the test data set. In general, the lack of consensus as to the 'best performing pipeline' illustrates the importance of the underlying dataset properties. Considering the dataset's characteristics under the operation, tweaking, and fine-tuning the settings of different pipelines may further, at least to some extent, diminish the variability in their accuracy.

7.2 | ITS rRNA

The nuclear ribosomal internal transcribed spacer (ITS) region is a standard marker in fungal metabarcoding studies (Nilsson et al., 2019). It is also taxonomically informative in other eukaryotic groups (e.g. flowering plants, mites, springtails; Anslan & Tedersoo, 2015; Banchi et al., 2020; Ben-David et al., 2007). The ITS region is highly variable in length among eukaryotic groups, complicating the bioinformatics analysis steps that rely on aligning (such as, e.g., mothur OTU clustering) or require uniform sequence length (such as, e.g., deblur). Pipelines such as NextITS, PIPITS (Gweon et al., 2015), and DAnIEL are developed explicitly for ITS amplicon analyses. Those pipelines implement the extraction of ITS sub-regions (ITS1/ITS2, or full ITS) to exclude flanking conservative regions (18S/5.8S/28S), which is optimal for taxonomic assignment accuracy (Bengtsson-Palme et al., 2013; Vu et al., 2022). SCATA is also optimized for the ITS region, and for other amplicon sequences which cannot be easily aligned. However, a few other universal pipelines, such as LotuS2, SEED 2, nf-core/ampliseq, PipeCraft2, MetaWorks, dadasnake, FROGS (all using ITSx; Bengtsson-Palme et al., 2013), and QIIME 2 (using the ITSxpress plugin; Rivers et al., 2018) incorporate the step for extracting the ITS sub-regions for optimal processing of ITS amplicon data. Because the ITS-subregions of some fungal groups may not sufficiently overlap during the paired-end data assembly

process, FROGS, PipeCraft2, Dadaist2 and Cascabel (latter two without ITSx) implement settings to also include nonassembled reads to ensure that taxa with longer ITS regions are not excluded (Bernard et al., 2021).

Although AMPtk, DADA2, eDNAflow, and gDAT were validated using ITS reads, these pipelines lack a step to clip the flanking regions from ITS reads. While ITS extraction tools may eliminate some fungal strains from the data, many false-positive molecular units are generated when this extraction process is excluded (Pauvert et al., 2019). To mitigate the detection of false-negatives, the exclusion of the ITS extraction may be more appropriate if the aim is to find specific target taxa, whereas the ITS extraction operation should be included in community ecology studies (Pauvert et al., 2019).

Tested on technical replicates from soil samples (i.e. DNA from the same sample sequenced twice), compositional matrices of ITS data from QIIME 2 and LotuS2 were more reproducible than native DADA2, where the latter did not incorporate an ITS extraction step (Özkurt et al., 2022). Differences in the ITS amplicon data analyses among various software (PipeCraft1, QIIME 2, PIPITS, LotuS1, and custom pipeline compiled on Galaxy platform) were evident also in the study by Anslan et al. (2018) where QIIME 2 and Galaxy-based pipelines did not include the ITS extraction step (because it was not yet implemented). Although the inclusion of ITS region extraction step lowers the amount of nontarget features, the latter study concluded that none of the tested workflows were able to fully filter out the erroneous sequences, which contributed to the demonstrated differences between pipelines.

7.3 | COI

Found in the mitochondria, the cytochrome oxidase subunit I (COI/CO1/cox1) is a standard animal barcode (Hajibabaei et al., 2011; Hebert et al., 2003). Compared with other suitable markers (e.g. mt 16S, ITS, 28S) for most metazoan groups, the reference database of the COI is vast (Porter & Hajibabaei, 2018) and COI fragments are extensively used in metabarcoding studies.

Metabarcoding of metazoan communities is increasingly employed in ecology, but the strategies for analysing the sequencing data vary largely across studies. Generally, the metabarcoding studies utilizing protein-coding genes (such as COI) have largely followed the bioinformatic workflows designed to characterize microbial diversity without adapting the workflows to the characteristics of protein-coding markers (Creedy et al., 2022). When processing protein-coding markers, the noise of nuclear mitochondrial pseudogenes (NUMTs) may inflate the richness estimates and thus introduce biases in biodiversity research using metabarcoding (Porter & Hajibabaei, 2021). Thus, the amino acid translation, but also the length of the read should be used to identify erroneous sequences (Creedy et al., 2022). Of the pipelines reviewed here, MetaWorks and VTAM implement a step of removing putative NUMTs, which alleviates the burden of manual curation of the features to produce more accurate richness estimates. The multimarker amplicon processing platform PipeCraft2 has also wrapped MetaWorks strategy of the pseudogene removal step. Apart from the full pipelines, the multisample features matrix may be processed with metaMATE (Andújar et al., 2021) to remove putative NUMTs and other erroneous sequences (based on, e.g., length and relative read abundance). Additionally, DARN (Zafeiropoulos et al., 2021), which makes use of the phylogenetic tree, aids in filtering out nontarget features and upon denoising, the characteristics of protein coding genes are also accounted for in the DnoisE (Antich et al., 2022). We will most likely see the latter module integrated into the already established pipelines in the near future.

7.4 | Other markers and multimarker pipelines

Besides the above-mentioned markers, other popular markers used for metabarcoding are mt 16S rRNA for Metazoa, mt 12S rRNA for fish (Miya et al., 2020), 18S rRNA for protists and other eukaryotes, 28S rRNA for nematodes and eukaryotes in general, rbcL for diatoms (Rimet et al., 2019), rbcL + matK and trnL for plants (CBOL Plant Working Group 1 et al., 2009; Taberlet et al., 2007), and 23S rRNA for photosynthetic microbes (Djemiel, Dequiedt, et al., 2020; Djemiel, Plassard, et al., 2020). A variety of pipelines have been applied for the analyses of the amplicon sequences from these markers. For example, MICCA, DADA2 for 18S rRNA (Harrison et al., 2021; Minerovic et al., 2020); DADA2, OBITools for mt 16S (Marquina et al., 2019; Thomsen & Sigsgaard, 2019); and custom built pipelines (using multiple third-party sequence data analysis tools) for other markers above (Anslan et al., 2021; Elbrecht et al., 2016; Liu & Zhang, 2021; Westfall et al., 2020). Benchmarked on mt 12S reads from both simulated and real eDNA data, the Barque pipeline demonstrated a small sensitivity improvement over QIIME 2 and OBITools (Mathon et al., 2021). Moreover, another VSEARCH-based custom pipeline found in the latter study, which was designed to match Barque's performance by adjusting the parameters and threshold, showed the same mean sensitivity as Barque, demonstrating that the careful choice of the tools for the required task provides accurate results.

Table 1 lists multimarker software that may be utilized for various markers. All of the developed application software contain the most crucial steps for basic metabarcoding data analyses, but the suitability of a software or workflow steps for a given marker should be assessed. For example, considering the length variability and alignability of the amplicon set is important when some pipeline steps (e.g. clustering) use alignment-based methods (such as in mothur) or require uniform read lengths (such as deblur denoising). When working, for example, rbcL amplicons (or amplicons from any other protein coding gene), validation is needed to ensure that the generated features do not represent potential pseudogenes (or off-target taxa) for biodiversity analyses. Some multimarker pipelines incorporate marker-specific steps, for example, extracting the ITS region, removing putative pseudogenes and off-target features (Appendix S1). Using a pipeline that is not restricted to a certain marker gene, but

where the above listed automated filtering processes are lacking, a manual feature curation step is usually required to filter out bioinformatically unfiltered noise or to validate that most of the noise has already been removed. Depending on the study context, different analytical pipelines may yield highly compatible results (e.g. Baltrušis et al., 2022; Kang et al., 2021), but the outcome and interpretation may also vary considerably (Anslan et al., 2018; Bailet et al., 2020; Pauvert et al., 2019; Straub et al., 2020) without the validation of the software suitability for a given marker.

8 | CONCLUDING REMARKS

The development of a wide range of metabarcoding data analysis pipelines illustrates the need for not only 'easy-to-use' software but also of specific customized workflows depending on the underlying sequencing data set. Although most of the precompiled pipelines largely mirror the functionalities of several software suites by incorporating steps from algorithms providing software suites, they offer easily executable automated alternatives for users with less bioinformatics experience. Additionally, many precompiled pipelines are supplemented with several possibilities for downstream analyses by wrapping various third-party tools. Applying different workflows on the same data will always demonstrate a certain level of variation among pipelines. These variations are usually most obvious in terms of the reported number of features. This generally derives from variations in filtering out spurious and low-abundant sequences (e.g. Edgar, 2017; Prodan et al., 2020). Therefore, one pipeline may produce a higher number of features per sample and the other much less, but the correlations between sample-wise richness from one to another result are in most cases very high (Baltrušis et al., 2022; Kang et al., 2021). However, depending on the analysed data set, this correlation pattern may be the opposite (Nearing et al., 2018) and pipeline settings should be carefully considered, especially when identifying rare taxa is imperative. Thus, although the automated pipelines have made the analyses easier and more reproducible, expertise is still required to validate the accuracy of the biological results. It is noteworthy that a pipeline's performance measured on mock community samples with relatively few species may vary when applied to a complex data set originating from environmental samples. Nevertheless, including a mock community control sample(s) in a study will certainly aid in identifying false positives and false negatives. A robust sense of the community patterns may be obtained by applying 'default' parameter values but fine-tuning of the parameters may be required to find an appropriate compromise between false positive removal and retention of true detections.

Table 1 and Figure 2 are aiming to provide assistance in narrowing down the desirable pipelines for the task. Once the potential target workhorses have been selected, one would naturally need to explore the respective user guides for more detailed information about the underlying procedures.

ACKNOWLEDGEMENTS

This work was supported by the European Regional Development Fund and the programme Mobilitas Pluss (MOBTP198). MH and TMP received funding from Genome Canada and Ontario Genomics through the Sequencing the Rivers for Environmental Assessment and Monitoring (STREAM) project. DS acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy, cluster of Excellence EXC2124 "Controlling microbes to fight infection" (CMFI), project ID 390838134. We wish to offer our heartfelt thanks to Sébastien Terrat, the leading developer of BIOCOM-PIPE (including ReClustOR) for the support extended to certain points in this tool. TV was supported by the Czech Science Foundation (21-17749S). We thank other FROGS'members Vincent Darbot, Lucas Auer and Olivier Rué, and Frédéric Mahé (swarm software author) for their fruitful exchanges on the ASV/OTU/cluster terminology. EC received funding through an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103449.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Not applicable.

ORCID

Ali Hakimzadeh https://orcid.org/0000-0003-1336-7445
Maria Bernard https://orcid.org/0000-0001-9005-5563
Dominik Buchner https://orcid.org/0000-0002-8499-5863
Christophe Djemiel https://orcid.org/0000-0002-5659-7876
Mikael Brandström Durling https://orcid.org/0000-0001-6485-197X
Vasco Elbrecht https://orcid.org/0000-0003-4672-7099
Mehrdad Hajibabaei https://orcid.org/0000-0002-8859-7977
Géraldine Pascal https://orcid.org/0000-0002-5250-6594
Martti Vasar https://orcid.org/0000-0002-4674-932X
Sten Anslan https://orcid.org/0000-0002-2299-454X

REFERENCES

Albanese, D., Fontana, P., de Filippo, C., Cavalieri, D., & Donati, C. (2015). MICCA: A complete and accurate software for taxonomic profiling of metagenomic data. *Scientific Reports*, *5*(1), 1–7. https://doi.org/10.1038/srep09743

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech, X. Z., Kightley, E. P., Thompson, L. R., Hyde, E. R., Gonzalez, A., & Knight, R. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems*, *2*(2), e00191-16. https://doi.org/10.1128/mSystems

Andújar, C., Creedy, T. J., Arribas, P., López, H., Salces-Castellano, A., Pérez-Delgado, A. J., Vogler, A. P., & Emerson, B. C. (2021). Validated removal of nuclear pseudogenes and sequencing artefacts from mitochondrial metabarcode data. *Molecular Ecology Resources*, *21*(6), 1772–1787. https://doi.org/10.1111/1755-0998.13337

Anslan, S., Bahram, M., Hiiesalu, I., & Tedersoo, L. (2017). PipeCraft: Flexible open-source toolkit for bioinformatics analysis of custom high-throughput amplicon sequencing data. *Molecular Ecology Resources*, *17*(6), e234–e240. https://doi.org/10.1111/1755-0998.12692

Anslan, S., Mikryukov, V., Armolaitis, K., Ankuda, J., Lazdina, D., Makovskis, K., Vesterdal, L., Schmidt, I. K., & Tedersoo, L. (2021). Highly comparable metabarcoding results from MGI-tech and Illumina sequencing platforms. *PeerJ*, *9*, e12254. https://doi.org/10.7717/peerj.12254

Anslan, S., Nilsson, R. H., Wurzbacher, C., Baldrian, P., Tedersoo, L., & Bahram, M. (2018). Great differences in performance and outcome of high-throughput sequencing data analysis platforms for fungal metabarcoding. *MycoKeys*, *39*, 29–40. https://doi.org/10.3897/mycokeys.39.28109

Anslan, S., & Tedersoo, L. (2015). Performance of cytochrome c oxidase subunit I (COI), ribosomal DNA large subunit (LSU) and internal transcribed spacer 2 (ITS2) in DNA barcoding of Collembola. *European Journal of Soil Biology*, *69*, 1–7. https://doi.org/10.1016/j.ejsobi.2015.04.001

Ansorge, R., Birolo, G., James, S. A., & Telatin, A. (2021). Dadaist2: A toolkit to automate and simplify statistical analysis and plotting of metabarcoding experiments. *International Journal of Molecular Sciences*, *22*(10), 5309. https://doi.org/10.3390/ijms22105309

Antich, A., Palacín, C., Turon, X., & Wangensteen, O. S. (2022). DnoisE: Distance denoising by entropy. An open-source parallelizable alternative for denoising sequence datasets. *PeerJ*, *10*, e12758. https://doi.org/10.7717/peerj.12758

Antich, A., Palacin, C., Wangensteen, O. S., & Turon, X. (2021). To denoise or to cluster, that is not the question: Optimizing pipelines for COI metabarcoding and metaphylogeography. *BMC Bioinformatics*, *22*(1), 177. https://doi.org/10.1186/s12859-021-04115-6

Asbun, A. A., Besseling, M. A., Balzano, S., van Bleijswijk, J. D. L., Witte, H. J., Villanueva, L., & Engelmann, J. C. (2020). Cascabel: A scalable and versatile amplicon sequence data analysis pipeline delivering reproducible and documented results. *Frontiers in Genetics*, *11*, 489357. https://doi.org/10.3389/fgene.2020.489357

Bai, J., Jhaney, I., & Wells, J. (2019). Developing a reproducible microbiome data analysis pipeline using the Amazon web services cloud for a cancer research group: Proof-of-concept study. *JMIR Medical Informatics*, *7*(4), e14667. https://doi.org/10.2196/14667

Bailet, B., Apothéloz-Perret-Gentil, L., Baričević, A., Chonova, T., Franc, A., Frigerio, J. M., Kelly, M., Mora, D., Pfannkuchen, M., Proft, S., Ramon, M., Vasselon, V., Zimmermann, J., & Kahlert, M. (2020). Diatom DNA metabarcoding for ecological assessment: Comparison among bioinformatics pipelines used in six European countries reveals the need for standardization. *Science of the Total Environment*, *745*, 140948. https://doi.org/10.1016/j.scitotenv.2020.140948

Baloğlu, B., Chen, Z., Elbrecht, V., Braukmann, T., MacDonald, S., & Steinke, D. (2021). A workflow for accurate metabarcoding using nanopore MinION sequencing. *Methods in Ecology and Evolution*, *12*(5), 794–804. https://doi.org/10.1111/2041-210X.13561

Baltrušis, P., Halvarsson, P., & Höglund, J. (2022). Estimation of the impact of three different bioinformatic pipelines on sheep nemabiome analysis. *Parasites & Vectors*, *15*(1), 1–12. https://doi.org/10.1186/s13071-022-05399-0

Banchi, E., Ametrano, C. G., Greco, S., Stanković, D., Muggia, L., & Pallavicini, A. (2020). PLANiTS: A curated sequence reference dataset for plant ITS DNA metabarcoding. *Database*, *2020*, baz155. https://doi.org/10.1093/database/baz155

Ben-David, T., Melamed, S., Gerson, U., & Morin, S. (2007). ITS2 sequences as barcodes for identifying and analyzing spider mites (Acari: Tetranychidae). *Experimental and Applied Acarology*, *41*(3), 169–181. https://doi.org/10.1186/s13071-022-05399-0

Bengtsson-Palme, J., Ryberg, M., Hartmann, M., Branco, S., Wang, Z., Godhe, A., de Wit, P., Sánchez-García, M., Ebersberger, I., de Sousa, F., Amend, A., & Nilsson, R. H. (2013). Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. *Methods in Ecology and Evolution*, *4*(10), 914–919. https://doi.org/10.1111/2041-210X.12073

Bernard, M., Rué, O., Mariadassou, M., & Pascal, G. (2021). FROGS: A powerful tool to analyse the diversity of fungi with special management of internal transcribed spacers. *Briefings in Bioinformatics*, *22*(6), bbab318. https://doi.org/10.1093/bib/bbab318

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., & Caporaso, J. G. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, *6*(1), 1–17. https://doi.org/10.1186/s40168-018-0470-z

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). Obitools: A unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, *16*(1), 176–182. https://doi.org/10.1111/1755-0998.12428

Brandt, M. I., Trouche, B., Quintric, L., Günther, B., Wincker, P., Poulain, J., & Arnaud-Haond, S. (2021). Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding. *Molecular Ecology Resources*, *21*(6), 1904–1921. https://doi.org/10.1111/1755-0998.13398

Brown, S. P., Veach, A. M., Rigdon-Huss, A. R., Grond, K., Lickteig, S. K., Lothamer, K., Oliver, A. K., & Jumpponen, A. (2015). Scraping the bottom of the barrel: Are rare high throughput sequences artifacts? *Fungal Ecology*, *13*, 221–225.

Bruce, K., Blackman, R. C., Bourlat, S. J., Hellström, M., Bakker, J., Bista, I., Bohmann, K., Bouchez, A., Brys, R., Clark, K., Elbrecht, V., & Deiner, K. (2021). *A Practical Guide to DNA-Based Methods for Biodiversity Assessment*. Pensoft Advanced Books. https://doi.org/10.3897/ab.e68634

Buchner, D., Macher, T.-H., & Leese, F. (2022). APSCALE: Advanced pipeline for simple yet comprehensive analyses of DNA metabarcoding data. *Bioinformatics*, *38*(20), 4817–4819. https://doi.org/10.1093/bioinformatics/btac588

Callahan, B. J., Grinevich, D., Thakur, S., Balamotis, M. A., & Yehezkel, T. B. (2021). Ultra-accurate microbial amplicon sequencing with synthetic long reads. *Microbiome*, *9*(1), 130. https://doi.org/10.1186/s40168-021-01072-3

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, *11*(12), 2639–2643. https://doi.org/10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Callahan, B. J., Wong, J., Heiner, C., Oh, S., Theriot, C. M., Gulati, A. S., McGill, S. K., & Dougherty, M. K. (2019). High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Research*, *47*(18), e103. https://doi.org/10.1093/nar/gkz569

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336. https://doi.org/10.1038/nmeth.f.303

Carlsen, T., Aas, A. B., Lindner, D., Vrålstad, T., Schumacher, T., & Kauserud, H. (2012). Don't make a mista (g) ke: Is tag switching an overlooked source of error in amplicon pyrosequencing studies? *Fungal Ecology*, *5*(6), 747–749.

Carøe, C., & Bohmann, K. (2020). Tagsteady: A metabarcoding library preparation protocol to avoid false assignment of sequences to samples. *Molecular Ecology Resources*, *20*(6), 1620–1631. https://doi.org/10.1111/1755-0998.13227

Castaño, C., Berlin, A., Brandström Durling, M., Ihrmark, K., Lindahl, B. D., Stenlid, J., Clemmensen, K. E., & Olson, Å. (2020). Optimized metabarcoding with Pacific biosciences enables semi-quantitative analysis of fungal communities. *New Phytologist*, *228*(3), 1149–1158. https://doi.org/10.1111/nph.16731

CBOL Plant Working Group 1, Hollingsworth, P. M., Hajibabaei, M., Ratnasingham, S., Chase, M., Cowan, R. S., Erickson, D. L., Fazekas, A. J., Graham, S. W., James, K. E., Kim, K.-J., Kress, W. J., Schneider, H., van Alphenstahl, J., Barrett, S. C. H., van den Berg, C., Bogarín, D., Burgess, K. S., Cameron, K. M., … Little, D. P. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, *106*(31), 12794–12797. https://doi.org/10.1073/pnas.0905845106

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Compson, Z. G., McClenaghan, B., Singer, G. A., Fahner, N. A., & Hajibabaei, M. (2020). Metabarcoding from microbes to mammals: Comprehensive bioassessment on a global scale. *Frontiers in Ecology and Evolution*, *8*, 581835. https://doi.org/10.3389/fevo.2020.581835

Copeland, M., Soh, J., Puca, A., Manning, M., & Gollob, D. (2015). Microsoft Azure and Cloud Computing. In *Microsoft azure* (pp. 3–26). Apress.

Couton, M., Baud, A., Daguin-Thiébaut, C., Corre, E., Comtet, T., & Viard, F. (2021). High-throughput sequencing on preservative ethanol is effective at jointly examining infraspecific and taxonomic diversity, although bioinformatics pipelines do not perform equally. *Ecology and Evolution*, *11*(10), 5533–5546. https://doi.org/10.1002/ece3.7453

Creedy, T. J., Andújar, C., Meramveliotakis, E., Noguerales, V., Overcast, I., Papadopoulou, A., Morlon, H., Vogler, A. P., Emerson, B. C., & Arribas, P. (2022). Coming of age for COI metabarcoding of whole organism community DNA: Towards bioinformatic harmonisation. *Molecular Ecology Resources*, *22*(3), 847–861. https://doi.org/10.1111/1755-0998.13502

Curd, E. E., Gold, Z., Kandlikar, G. S., Gomer, J., Ogden, M., O'Connell, T., Pipes, L., Schweizer, T. M., Rabichow, L., Lin, M., Shi, B., & Meyer, R. S. (2019). Anacapa toolkit: An environmental DNA toolkit for processing multilocus metabarcode datasets. *Methods in Ecology and Evolution*, *10*(9), 1469–1475. https://doi.org/10.1111/2041-210X.13214

de Santiago, A., Pereira, T. J., Mincks, S. L., & Bik, H. M. (202). Dataset complexity impacts both MOTU delimitation and biodiversity estimates in eukaryotic 18S rRNA metabarcoding studies. *Environmental DNA*, *4*(2), 363–384. https://doi.org/10.1002/edn3.255

di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. https://doi.org/10.1038/nbt.3820

Djemiel, C., Dequiedt, S., Karimi, B., Cottin, A., Girier, T., El Djoudi, Y., Wincker, P., Lelièvre, M., Mondy, S., Chemidlin Prévost-Bouré, N., Maron, P.-A., Ranjard, L., & Terrat, S. (2020). BIOCOM-PIPE: A new user-friendly metabarcoding pipeline for the characterization of microbial diversity from 16S, 18S and 23S rRNA gene amplicons. *BMC Bioinformatics*, *21*(1), 492. https://doi.org/10.1186/s12859-020-03829-3

Djemiel, C., Plassard, D., Terrat, S., Crouzet, O., Sauze, J., Mondy, S., Nowak, V., Wingate, L., Ogée, J., & Maron, P. A. (2020). μgreen-db: A reference database for the 23S rRNA gene of eukaryotic plastids and cyanobacteria. *Scientific Reports*, *10*(1), 1–11. https://doi.org/10.1038/s41598-020-62555-1

Durling, M. B., Clemmensen, K. E., Stenlid, J., & Lindahl, B. (2011). *SCATA-an efficient bioinformatic pipeline for species identification and quantification after high-throughput sequencing of tagged amplicons*. https://scata.mykopat.slu.se/

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Edgar, R. C. (2016a). SINTAX: A simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv*, 74161. https://doi.org/10.1101/074161

Edgar, R. C. (2016b). UNOISE2: Improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv*, *81257*. https://doi.org/10.1101/081257

Edgar, R. C. (2017). Accuracy of microbial community diversity estimated by closed-and open-reference OTUs. *PeerJ*, *5*, e3889. https://doi.org/10.7717/peerj.3889

Edgar, R. C. (2018a). Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ*, *6*, e4652. https://doi.org/10.7717/peerj.4652

Edgar, R. C. (2018b). UNCROSS2: Identification of cross-talk in 16S rRNA OTU tables. *bioRxiv*, 400762. https://doi.org/10.1101/400762

Edgar, R. C., & Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics*, *31*(21), 3476–3482. https://doi.org/10.1093/bioinformatics/btv401

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, *27*(16), 2194–2200. https://doi.org/10.1093/bioinformatics/btr381

Elbrecht, V., Taberlet, P., Dejean, T., Valentini, A., Usseglio-Polatera, P., Beisel, J. N., Coissac, E., Boyer, F., & Leese, F. (2016). Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. *PeerJ*, *4*, e1966. https://doi.org/10.7717/peerj.1966

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, *38*, 276–278. https://doi.org/10.1038/s41587-020-0439-x

Escudié, F., Auer, L., Bernard, M., Mariadassou, M., Cauquil, L., Vidal, K., Maman, S., Hernandez-Raquet, G., Combes, S., & Pascal, G. (2018). FROGS: Find, rapidly, OTUs with galaxy solution. *Bioinformatics*, *34*(8), 1287–1294. https://doi.org/10.1093/bioinformatics/btx791

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications*, *8*(1), 1–11. https://doi.org/10.1038/s41467-017-01312-x

Furneaux, B., Bahram, M., Rosling, A., Yorou, N. S., & Ryberg, M. (2021). Long-and short-read metabarcoding technologies reveal similar spatiotemporal structures in fungal communities.

Molecular Ecology Resources, *21*(6), 1833–1849. https://doi.org/10.1111/1755-0998.13387

Galaxy Community. (2022). The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research*, *50*(W1), W345–W351. https://doi.org/10.1093/nar/gkac247

Glassman, S. I., & Martiny, J. B. (2018). Broadscale ecological patterns are robust to use of exact sequence variants versus operational taxonomic units. *MSphere*, *3*(4), e00148-18. https://doi.org/10.1128/mSphere.00148-18

Gold, Z., Curd, E. E., Goodwin, K. D., Choi, E. S., Frable, B. W., Thompson, A. R., Walker, H. J., Jr., Burton, R. S., Kacev, D., Martz, L. D., & Barber, P. H. (2021). Improving metabarcoding taxonomic assignment: A case study of fishes in a large marine ecosystem. *Molecular Ecology Resources*, *21*(7), 2546–2564. https://doi.org/10.1111/1755-0998.13450

González, A., Dubut, V., Corse, E., Mekdad, R., Dechatre, T., Castet, U., Hebert, R., & Meglécz, E. (2023). VTAM: A robust pipeline for validating metabarcoding data using controls. *Computational and Structural Biotechnology Journal.*, *21*, 1151–1156. https://doi.org/10.1016/j.csbj.2023.01.034

Gweon, H. S., Oliver, A., Taylor, J., Booth, T., Gibbs, M., Read, D. S., Griffiths, R. I., & Schonrogge, K. (2015). PIPITS: An automated pipeline for analyses of fungal internal transcribed spacer sequences from the Illumina sequencing platform. *Methods in Ecology and Evolution/British Ecological Society*, *6*(8), 973–980. https://doi.org/10.1111/2041-210X.12399

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A., & Baird, D. J. (2011). Environmental barcoding: A next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS One*, *6*(4), e17497. https://doi.org/10.1371/journal.pone.0017497

Harrison, J. P., Chronopoulou, P. M., Salonen, I. S., Jilbert, T., & Koho, K. A. (2021). 16S and 18S rRNA gene metabarcoding provide congruent information on the responses of sediment communities to eutrophication. *Frontiers in Marine Science*, *8*, 708716. https://doi.org/10.3389/fmars.2021.708716

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*, 313–321. https://doi.org/10.1098/rspb.2002.2218

Heeger, F., Bourne, E. C., Baschien, C., Yurkov, A., Bunk, B., Spröer, C., Overmann, J., Mazzoni, C. J., & Monaghan, M. T. (2018). Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments. *Molecular Ecology Resources*, *18*(6), 1500–1514. https://doi.org/10.1111/1755-0998.12937

Hildebrand, F., Tadeo, R., Voigt, A. Y., Bork, P., & Raes, J. (2014). LotuS: An efficient and user-friendly OTU processing pipeline. *Microbiome*, *2*(1), 1–7. https://doi.org/10.1186/2049-2618-2-30

Hleap, J. S., Littlefair, J. E., Steinke, D., Hebert, P. D., & Cristescu, M. E. (2021). Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes. *Molecular Ecology Resources*, *21*(7), 2190–2203. https://doi.org/10.1111/1755-0998.13407

Hupfauf, S., Etemadi, M., Juárez, M. F.-D., Gómez-Brandón, M., Insam, H., & Podmirseg, S. M. (2020). CoMA – An intuitive and user-friendly pipeline for amplicon-sequencing data analysis. *PLoS One*, *15*(12), e0243241. https://doi.org/10.1371/journal.pone.0243241

Huse, S. M., Welch, D. M., Morrison, H. G., & Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology*, *12*(7), 1889–1898. https://doi.org/10.1111/j.1462-2920.2010.02193.x

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, *17*(3), 377–386. https://doi.org/10.1101/gr.5969107

Hussain, A., & Aleem, M. (2018). GoCJ: Google cloud jobs dataset for distributed and cloud computing infrastructures. *Data*, *3*(4), 38. https://doi.org/10.3390/data3040038

Kaehler, B. D., Bokulich, N. A., McDonald, D., Knight, R., Caporaso, J. G., & Huttley, G. A. (2019). Species abundance information improves sequence taxonomy classification accuracy. *Nature Communications*, *10*(1), 4643. https://doi.org/10.1038/s41467-019-12669-6

Kang, W., Anslan, S., Börner, N., Schwarz, A., Schmidt, R., Künzel, S., Rioual, P., Echeverría-Galindo, P., Vences, M., Wang, J., & Schwalb, A. (2021). Diatom metabarcoding and microscopic analyses from sediment samples at Lake Nam Co, Tibet: The effect of sample-size and bioinformatics on the identified communities. *Ecological Indicators*, *121*, 107070. https://doi.org/10.1016/j.ecolind.2020.107070

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L.-I., McDonald, D., Melnik, A. V., Morton, J. T., Navas, J., Quinn, R. A., Sanders, J. G., Swafford, A. D., Thompson, L. R., Tripathi, A., Xu, Z. Z., … Dorrestein, P. C. (2018). Best practices for analysing microbiomes. *Nature Reviews. Microbiology*, *16*(7), 410–422. https://doi.org/10.1038/s41579-018-0029-9

Koster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. *PLoS One*, *12*(5), e0177459. https://doi.org/10.1371/journal.pone.0177459

Laehnemann, D., Borkhardt, A., & McHardy, A. C. (2016). Denoising DNA deep sequencing data—High-throughput sequencing errors and their correction. *Briefings in Bioinformatics*, *17*(1), 154–179. https://doi.org/10.1093/bib/bbv029

Lear, G., Dickie, I., Banks, J., Boyer, S., Buckley, H. L., Buckley, T. R., Buckley, T. R., Cruickshank, R., & Holdaway, R. (2018). Methods for the extraction, storage, amplification and sequencing of DNA from environmental samples. *New Zealand Journal of Ecology*, *42*(1), 10–50A. https://doi.org/10.20417/nzjecol.42.9

Lindgreen, S. (2012). AdapterRemoval: Easy cleaning of next-generation sequencing reads. *BMC Research Notes*, *5*(1), 1–7. https://doi.org/10.1186/1756-0500-5-337

Liu, J., & Zhang, H. (2021). Combining multiple markers in environmental DNA metabarcoding to assess deep-sea benthic biodiversity. *Frontiers in Marine Science*, *8*, 684955. https://doi.org/10.3389/fmars.2021.684955

Loos, D., Zhang, L., Beemelmanns, C., Kurzai, O., & Panagiotou, G. (2021). DAnIEL: A user-friendly web server for fungal ITS amplicon sequencing data. *Frontiers in Microbiology*, *12*, 720513. https://doi.org/10.3389/fmicb.2021.720513

Mahé, F., Czech, L., Stamatakis, A., Quince, C., de Vargas, C., Dunthorn, M., & Rognes, T. (2022). Swarm v3: Towards tera-scale amplicon clustering. *Bioinformatics*, *38*(1), 267–269. https://doi.org/10.1093/bioinformatics/btab493

Marquina, D., Esparza-Salas, R., Roslin, T., & Ronquist, F. (2019). Establishing arthropod community composition using metabarcoding: Surprising inconsistencies between soil samples and preservative ethanol and homogenate from malaise trap catches. *Molecular Ecology Resources*, *19*(6), 1516–1530. https://doi.org/10.1111/1755-0998.13071

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, *17*(1), 10–12. https://doi.org/10.14806/ej.17.1.200

Mathon, L., Valentini, A., Guérin, P. E., Normandeau, E., Noel, C., Lionnet, C., Boulanger, E., Thuiller, W., Bernatchez, L., Mouillot, D., Dejean, T., & Manel, S. (2021). Benchmarking bioinformatic tools for fast and accurate eDNA metabarcoding species identification. *Molecular Ecology Resources*, *21*(7), 2565–2579. https://doi.org/10.1111/1755-0998.13430

McGee, K. M., Robinson, C. V., & Hajibabaei, M. (2019). Gaps in DNA-based biomonitoring across the globe. *Frontiers in Ecology and Evolution*, *7*, 337. https://doi.org/10.3389/fevo.2019.00337

Mikryukov, V., Anslan, S., & Tedersoo, L. (2022). *NextITS: A pipeline for metabarcoding fungi and other eukaryotes with full-length ITS sequenced with PacBio*. https://github.com/vmikk/NextITS

Minerovic, A. D., Potapova, M. G., Sales, C. M., Price, J. R., & Enache, M. D. (2020). 18S-V9 DNA metabarcoding detects the effect of water-quality impairment on stream biofilm eukaryotic assemblages. *Ecological Indicators*, *113*, 106225. https://doi.org/10.1016/j.ecolind.2020.106225

Miya, M., Gotoh, R. O., & Sado, T. (2020). MiFish metabarcoding: A high-throughput approach for simultaneous detection of multiple fish species from environmental DNA and other samples. *Fisheries Science*, *86*(6), 939–970. https://doi.org/10.1007/s12562-020-01461-x

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). Sustainable data analysis with Snakemake. *F1000Research*, *10*, 33. https://doi.org/10.12688/f1000research.29032.2

Mousavi-Derazmahalleh, M., Stott, A., Lines, R., Peverley, G., Nester, G., Simpson, T., Zawierta, M., De la Pierre, M., Bunce, M., & Christophersen, C. T. (2021). eDNAFlow, an automated, reproducible and scalable workflow for analysis of environmental DNA sequences exploiting Nextflow and singularity. *Molecular Ecology Resources*, *21*(5), 1697–1704. https://doi.org/10.1111/1755-0998.13356

Nearing, J. T., Douglas, G. M., Comeau, A. M., & Langille, M. G. (2018). Denoising the Denoisers: An independent evaluation of microbiome sequence error-correction approaches. *PeerJ*, *6*, e5364. https://doi.org/10.7717/peerj.5364

Nilsson, R. H., Anslan, S., Bahram, M., Wurzbacher, C., Baldrian, P., & Tedersoo, L. (2019). Mycobiome diversity: High-throughput sequencing and identification of fungi. *Nature Reviews. Microbiology*, *17*(2), 95–109. https://doi.org/10.1038/s41579-018-0116-y

Nilsson, R. H., Wurzbacher, C., Bahram, M., Coimbra, V. R., Larsson, E., Tedersoo, L., Eriksson, J., Ritter, C. D., Svantesson, S., Sánchez-García, M., Ryberg, M., & Abarenkov, K. (2016). Top 50 most wanted fungi. *MycoKeys*, *12*, 29–40. https://doi.org/10.3897/mycokeys.12.7553

Özkurt, E., Fritscher, J., Soranzo, N., Ng, D. Y. K., Davey, R. P., Bahram, M., & Hildebrand, F. (2022). LotuS2: An ultrafast and highly accurate tool for amplicon sequencing analysis. *Microbiome*, *10*(1), 176. https://doi.org/10.1186/s40168-022-01365-1

Palmer, J. M., Jusino, M. A., Banik, M. T., & Lindner, D. L. (2018). Non-biological synthetic spike-in controls and the AMPtk software pipeline improve mycobiome data. *PeerJ*, *6*, e4925. https://doi.org/10.7717/peerj.4925

Pauvert, C., Buee, M., Laval, V., Edel-Hermann, V., Fauchery, L., Gautier, A., Lesur, I., Vallance, J., & Vacher, C. (2019). Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline. *Fungal Ecology*, *41*, 23–33. https://doi.org/10.1016/j.funeco.2019.03.005

Pollock, J., Glendinning, L., Wisedchanwet, T., & Watson, M. (2018). The madness of microbiome: Attempting to find consensus "best practice" for 16S microbiome studies. *Applied and Environmental Microbiology*, *84*(7), e02627-17. https://doi.org/10.1128/AEM.02627-17

Porter, T. M., & Hajibabaei, M. (2018). Automated high throughput animal CO1 metabarcode classification. *Scientific Reports*, *8*(1), 4226. https://doi.org/10.1038/s41598-018-22505-4

Porter, T. M., & Hajibabaei, M. (2020). Putting COI metabarcoding in context: The utility of exact sequence variants (ESVs) in biodiversity analysis. *Frontiers in Ecology and Evolution*, *8*, 248. https://doi.org/10.3389/fevo.2020.00248

Porter, T. M., & Hajibabaei, M. (2021). Profile hidden Markov model sequence analysis can help remove putative pseudogenes from DNA barcoding and metabarcoding datasets. *BMC Bioinformatics*, *22*(1), 1–20. https://doi.org/10.1186/s12859-021-04180-x

Porter, T. M., & Hajibabaei, M. (2022). MetaWorks: A flexible, scalable bioinformatic pipeline for high-throughput multi-marker biodiversity assessments. *PLoS One*, *17*(9), e0274260. https://doi.org/10.1371/journal.pone.0274260

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M., & Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS One*, *15*(1), e0227434. https://doi.org/10.1371/journal.pone.0227434

Reeder, J., & Knight, R. (2009). The'rare biosphere': A reality check. *Nature Methods*, *6*(9), 636–637. https://doi.org/10.1038/nmeth0909-636

Reitmeier, S., Hitch, T. C., Treichel, N., Fikas, N., Hausmann, B., Ramer-Tait, A. E., Neuhaus, K., Berry, D., Haller, D., Lagkouvardos, I., & Clavel, T. (2021). Handling of spurious sequences affects the outcome of high-throughput 16S rRNA gene amplicon profiling. *ISME Communications*, *1*(1), 1–12. https://doi.org/10.1038/s43705-021-00033-z

Richardson, R. T., Bengtsson-Palme, J., & Johnson, R. M. (2017). Evaluating and optimizing the performance of software commonly used for the taxonomic classification of DNA metabarcoding sequence data. *Molecular Ecology Resources*, *17*(4), 760–769. https://doi.org/10.1111/1755-0998.12628

Rimet, F., Gusev, E., Kahlert, M., Kelly, M. G., Kulikovskiy, M., Maltsev, Y., Mann, D. G., Pfannkuchen, M., Trobajo, R., Vasselon, V., Zimmermann, J., & Bouchez, A. (2019). Diat. Barcode, an open-access curated barcode library for diatoms. *Scientific Reports*, *9*(1), 15116. https://doi.org/10.1038/s41598-019-51500-6

Rivers, A. R., Weber, K. C., Gardner, T. G., Liu, S., & Armstrong, S. D. (2018). ITSxpress: Software to rapidly trim internally transcribed spacer sequences with quality scores for marker gene analysis. *F1000Research*, *7*, 7. https://doi.org/10.12688/f1000research.15704.1

Rodriguez-Martinez, S., Klaminder, J., Morlock, M. A., Dalén, L., & Huang, D. T. (2022). The topological nature of tag jumping in environmental DNA metabarcoding studies. *Molecular Ecology Resources.*, *23*, 621–631. https://doi.org/10.1111/1755-0998.13745

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Rosen, G. L., Reichenberger, E. R., & Rosenfeld, A. M. (2011). NBC: The naive Bayes classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics*, *27*(1), 127–129. https://doi.org/10.1093/bioinformatics/btq619

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D., & Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, *15*(6), 1289–1303. https://doi.org/10.1111/1755-0998.12402

Singer, G. A. C., Fahner, N. A., Barnes, J. G., McCarthy, A., & Hajibabaei, M. (2019). Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: A case study of eDNA metabarcoding seawater. *Scientific Reports*, *9*(1), 5991. https://doi.org/10.1038/s41598-019-42455-9

Song, H., Buhay, J. E., Whiting, M. F., & Crandall, K. A. (2008). Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences of the United*

States of America, *105*(36), 13486–13491. https://doi.org/10.1073/pnas.0803076105

Staats, M., Arulandhu, A. J., Gravendeel, B., Holst-Jensen, A., Scholtens, I., Peelen, T., Prins, T. W., & Kok, E. (2016). Advances in DNA metabarcoding for food and wildlife forensic species identification. *Analytical and Bioanalytical Chemistry*, *408*(17), 4615–4630. https://doi.org/10.1007/s00216-016-9595-8

Straub, D., Blackwell, N., Langarica-Fuentes, A., Peltzer, A., Nahnsen, S., & Kleindienst, S. (2020). Interpretations of environmental microbial community studies are biased by the selected 16S rRNA (gene) amplicon sequencing pipeline. *Frontiers in Microbiology*, *11*, 550420. https://doi.org/10.3389/fmicb.2020.550420

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA: For biodiversity research and monitoring*. Oxford University Press. https://doi.org/10.1093/oso/9780198767220.001.0001

Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental dna. *Molecular Ecology*, *21*(8), 1789–1793. https://doi.org/10.1111/j.1365-294X.2012.05542.x

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*(8), 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., Vermat, T., Corthier, G., Brochmann, C., & Willerslev, E. (2007). Power and limitations of the chloroplast trn L (UAA) intron for plant DNA barcoding. *Nucleic Acids Research*, *35*(3), e14. https://doi.org/10.1093/nar/gkl938

Tedersoo, L., Albertsen, M., Anslan, S., & Callahan, B. (2021). Perspectives and benefits of high-throughput long-read sequencing in microbial ecology. *Applied and Environmental Microbiology*, *87*(17), e00626-21. https://doi.org/10.1128/AEM.00626-21

Tedersoo, L., & Anslan, S. (2019). Towards PacBio-based pan-eukaryote metabarcoding using full-length ITS sequences. *Environmental Microbiology Reports*, *11*(5), 659–668. https://doi.org/10.1111/1758-2229.12776

Tedersoo, L., Bahram, M., Zinger, L., Nilsson, R. H., Kennedy, P. G., Yang, T., Anslan, S., & Mikryukov, V. (2022). Best practices in metabarcoding of fungi: From experimental design to results. *Molecular Ecology*, *31*(10), 2769–2795. https://doi.org/10.1111/mec.16460

Terrat, S., Djemiel, C., Journay, C., Karimi, B., Dequiedt, S., Horrigue, W., Maron, P. A., Chemidlin Prévost-Bouré, N., & Ranjard, L. (2020). ReClustOR: A re-clustering tool using an open-reference method that improves operational taxonomic unit definition. *Methods in Ecology and Evolution*, *11*(1), 168–180. https://doi.org/10.1111/2041-210X.13316

Thompson, L. R., Anderson, S. R., den Uyl, P. A., Patin, N. V., Lim, S. J., Sanderson, G., & Goodwin, K. D. (2022). Tourmaline: A containerized workflow for rapid and iterable amplicon sequence analysis using QIIME 2 and Snakemake. *GigaScience*, *11*, giac066. https://doi.org/10.1093/gigascience/giac066

Thomsen, P. F., & Sigsgaard, E. E. (2019). Environmental DNA metabarcoding of wild flowers reveals diverse communities of terrestrial arthropods. *Ecology and Evolution*, *9*(4), 1665–1679. https://doi.org/10.1002/ece3.4809

Vasar, M., Davison, J., Neuenkamp, L., Sepp, S.-K., Young, J. P. W., Moora, M., & Öpik, M. (2021). User-friendly bioinformatics pipeline gDAT (graphical downstream analysis tool) for analysing rDNA sequences. *Molecular Ecology Resources*, *21*(4), 1380–1392. https://doi.org/10.1111/1755-0998.13340

Vetrovský, T., Baldrian, P., & Morais, D. (2018). SEED 2: A user-friendly platform for amplicon high-throughput sequencing data analyses. *Bioinformatics*, *34*(13), 2292–2294. https://doi.org/10.1093/bioinformatics/bty071

Vu, D., Nilsson, R. H., & Verkley, G. J. M. (2022). Dnabarcoder: An open-source software package for analysing and predicting DNA sequence similarity cutoffs for fungal sequence identification. *Molecular Ecology Resources*., *22*, 2793–2809. https://doi.org/10.1111/1755-0998.13651

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267. https://doi.org/10.1128/AEM.00062-07

Weißbecker, C., Schnabel, B., & Heintz-Buschart, A. (2020). Dadasnake, a Snakemake implementation of DADA2 to process amplicon sequencing data for microbial ecology. *GigaScience*, *9*(12), giaa135. https://doi.org/10.1093/gigascience/giaa135

Weigand, H., Beermann, A. J., Čiampor, F., Costa, F. O., Csabai, Z., Duarte, S., Geiger, M. F., Grabowski, M., Rimet, F., Rulik, B., Strand, M., Szucsich, N., Weigand, A. M., Willassen, E., Wyler, S. A., Bouchez, A., Borja, A., Čiamporová-Zaťovičová, Z., Ferreira, S., … Ekrem, T. (2019). DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. *Science of the Total Environment*, *678*, 499–524. https://doi.org/10.1016/j.scitotenv.2019.04.247

Westfall, K. M., Therriault, T. W., & Abbott, C. L. (2020). A new approach to molecular biosurveillance of invasive species using DNA metabarcoding. *Global Change Biology*, *26*(2), 1012–1022. https://doi.org/10.1111/gcb.14886

Wratten, L., Wilm, A., & Göke, J. (2021). Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature Methods*, *18*(10), 1161–1168. https://doi.org/10.1038/s41592-021-01254-9

Zafeiropoulos, H., Gargan, L., Hintikka, S., Pavloudi, C., & Carlsson, J. (2021). The dark mAtteR iNvestigator (DARN) tool: Getting to know the known unknowns in COI amplicon data. *Metabarcoding and Metagenomics*, *5*, e69657. https://doi.org/10.3897/mbmg.5.69657

Zafeiropoulos, H., Viet, H. Q., Vasileiadou, K., Potirakis, A., Arvanitidis, C., Topalis, P., Pavloudi, C., & Pafilis, E. (2020). PEMA: A flexible pipeline for environmental DNA Metabarcoding analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes. *GigaScience*, *9*(3), giaa022. https://doi.org/10.1093/gigascience/giaa022

Zinger, L., Lionnet, C., Benoiston, A. S., Donald, J., Mercier, C., & Boyer, F. (2021). metabaR: An R package for the evaluation and improvement of DNA metabarcoding data quality. *Methods in Ecology and Evolution*, *12*(4), 586–592. https://doi.org/10.1111/2041-210X.13552

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hakimzadeh, A., Abdala Asbun, A., Albanese, D., Bernard, M., Buchner, D., Callahan, B., Caporaso, J. G., Curd, E., Djemiel, C., Brandström Durling, M., Elbrecht, V., Gold, Z., Gweon, H. S., Hajibabaei, M., Hildebrand, F., Mikryukov, V., Normandeau, E., Özkurt, E., M. Palmer, J. … Anslan, S. (2023). A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses. *Molecular Ecology Resources*, *00*, 1–17. https://doi.org/10.1111/1755-0998.13847

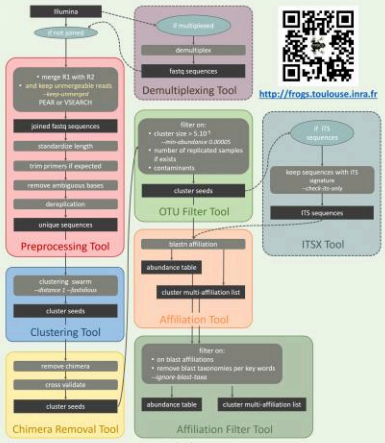
FROGSFUNC:
Smart integration of PICRUSt2 software into FROGS pipeline

Vincent DARBOT[1], Moussa SAMB[1], Maria BERNARD[2], Olivier RUÉ[3] and Géraldine PASCAL[1]

[1] GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France. [2] Univ. Paris-Saclay, INRAE, AgroParisTech, GABI, SIGENAE, F-78352, Jouy-en-Josas, France. [3] Univ. Paris-Saclay, INRAE, BioinfOmics, MIGALE bioinformatics facility, Jouy-en-Josas, France

Corresponding Author: vincent.darbot@inrae.fr

Background



Metabarcoding is the large-scale taxonomic identification of complex environmental samples via analysis of DNA reads of one marker gene (16S, ITS, 18S, COI, etc.).

The aim of metabarcoding analysis is to provide a table of abundance of OTUs/ASVs, as close as possible to the species, per sample as well as a descriptive statistical analysis of the composition of the targeted microbial population of the samples.

The various tools developed within FROGS [1,2] offers such results. They allow users to process their data in command lines or *via* in a user-friendly Galaxy[3] interface and to obtain different graphical and descriptive outputs.

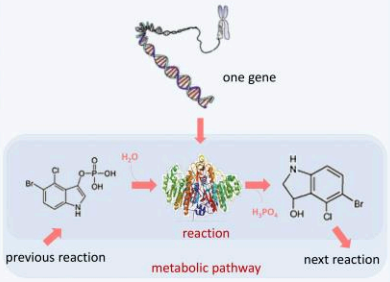
Unlike metagenomics[4], metabarcoding does not provide these functional profiles of a microbial population, by being restricted to one marker gene.

PICRUSt2 [5] bypass this restriction and obtain a prediction of the functional potential of a sample, at low cost.

Firstly, PICRUSt2 placed the marker gene (16S, ITS or 18S) sequences of interest into a its reference tree, that is used as the basis of functional predictions. After, it predicts number of marker and function copy number in each OTU. Then, for each sample, it calculates functions abundances and finally, pathway abundances are inferred, based on functional profile.

What metabolic functions are presents in a microbial community ?

Functional inference: we assume that an organism has a metabolic function by the presence in the organism's genome of a known sequence having that function.



Functional profiling within FROGS

Workflow

PICRUSt2 placed the marker gene sequences (16S, ITS or 18S) of interest into a its reference tree, that is used as the basis of functional predictions

Insert sequences

FROGSFUNC step1 placeseqs tool

Running hidden-state prediction to predict function abundances with castor-R of each OTUs placed in the PICRUSt2 reference phylogenetic tree

Predicted gene copies number per sequence

FROGSFUNC step2 copy numbers tool

Prediction of functions weighted by the relative abundance of OTUs in the community.

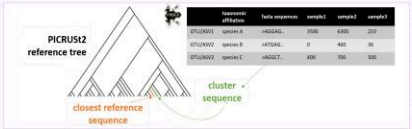
Function abundance predictions per sample

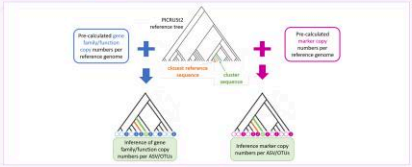
FROGSFUNC step3 functions tool

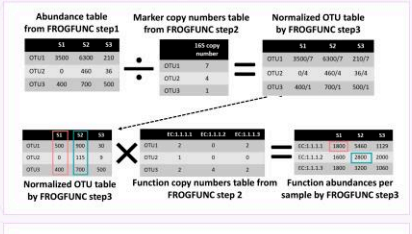
Inferrance of MetaCyc/KEGG pathway abundances based on function number abundances.

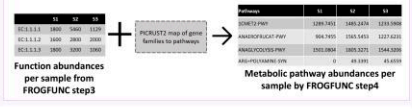
Pathway abundance predictions per sample

FROGSFUNC step4 pathways tool

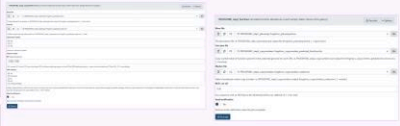






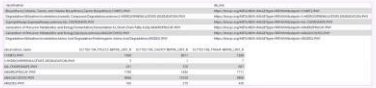


Ability to use Galaxy interfaces



Various tables and graphical outputs are displayed to make the experience intuitive

FROGS highlights important information that are very useful for data interpretation. Some "hidden" PICRUSt2 outputs are exploited: reporting incongruence between taxonomic affiliations, Nearest Sequenced Taxon Index (NSTI) threshold confidence indicator, decision support graphic to help choosing the NSTI threshold



References

[1] Escudié F, Auer L, Bernard M, Mariadassou M, Cauquil L, Vidal K, Maman S, Hernandez-Raquet G, Combes S, Pascal G. FROGS: Find, Rapidly, OTUs with Galaxy Solution. Bioinformatics. 2018 Apr 15;34(8):1287-1294. doi: 10.1093/bioinformatics/btx791. PMID: 29228191.
[2] Bernard M, Rué, O, Mariadassou M, and Pascal G; FROGS: a powerful tool to analyse the diversity of fungi with special management of internal transcribed spacers. Briefings in Bioinformatics. 2021, Nov. doi : 10.1093/bib/bbab318
[3] Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018 Jul 2;46(W1):W537-W544. doi: 10.1093/nar/gky379. PMID: 29790989; PMCID: PMC6030816.
[4] Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. Microb Inform Exp. 2012 Feb 9;2(1):3. doi: 10.1186/2042-5783-2-3. PMID: 22587947; PMCID: PMC3351745.
[5] Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, Huttenhower C, Langille MGI. PICRUSt2 for prediction of metagenome functions. Nat Biotechnol. 2020 Jun;38(6):685-688. doi: 10.1038/s41587-020-0548-6. PMID: 32483366; PMCID: PMC7365738.

Vincent Darbot's work has been supported by RESALAB OUEST

Poster présenté par Vincent Darbot lors de la conférence JOBIM 2022.

J.2. Liste des figures annexes

Vous trouverez dans cette annexe des figures additionnels référencés dans le manuscrit.

- La **Figure annexe J.2-1** et la **Figure annexe J.2-2** sont des résultats additionnels de l'analyse triadique partielle présentée dans le paragraphe **F.3.2**. Elles correspondent aux visualisations des structures de données des métagénomes, des métatranscriptomes et de leur compromis en tenant compte d'un *a priori* sur l'effet de leur appartenance à un groupe « lignée x régime » ou au contraire en ignorant leur appartenance à un groupe « lignée x régime ».

- La **Figure annexe J.2-3** permet la visualisation des fonctions différentiellement abondantes du métabolisme du méthane détectées sur l'analyse de métabarcoding comparativement aux fonctions différentiellement exprimées décrite dans le paragraphe **F.4.2**.

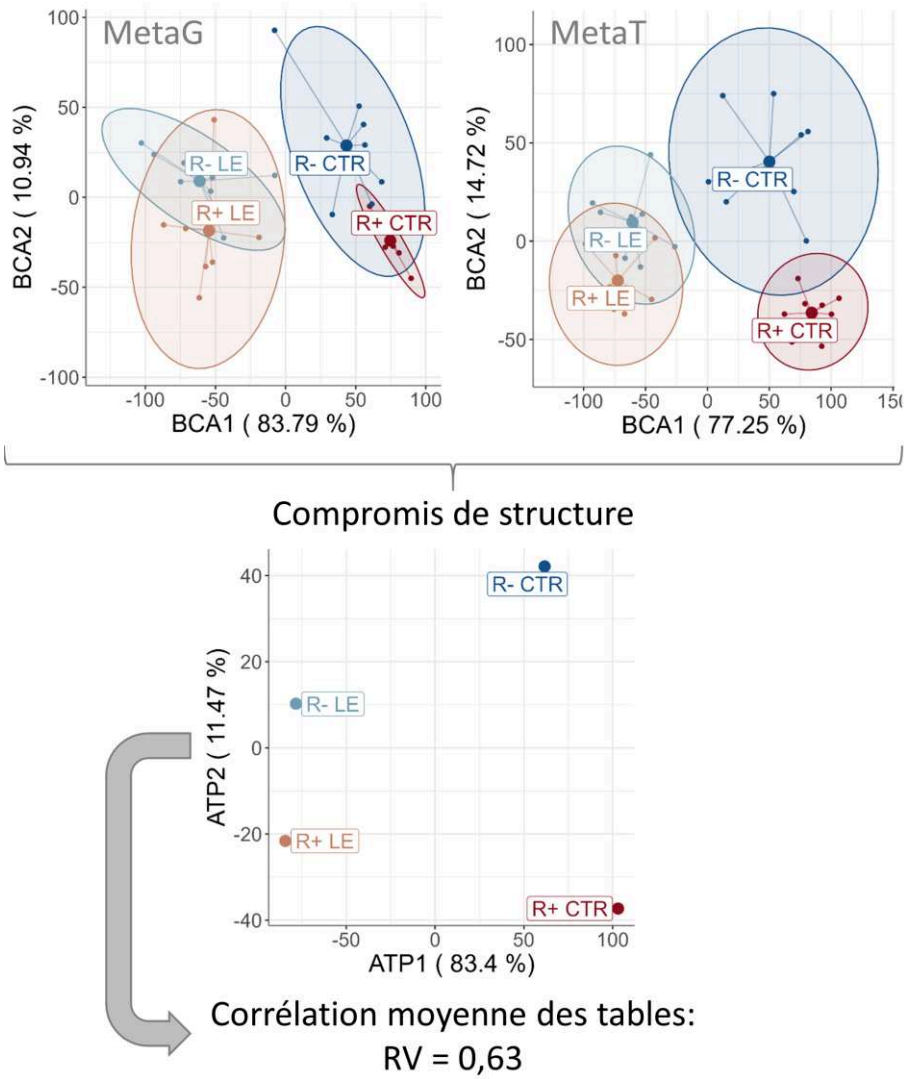


Figure annexe J.2-1 : Analyse triadique partielle sur des analyses entre groupe « lignée x régime » (BCA) ; L'analyse BCA confirme un effet significatif de l'appartenance à un groupe (*P*-value < 0,001 d'un test par 9 999 permutations des fonctions).

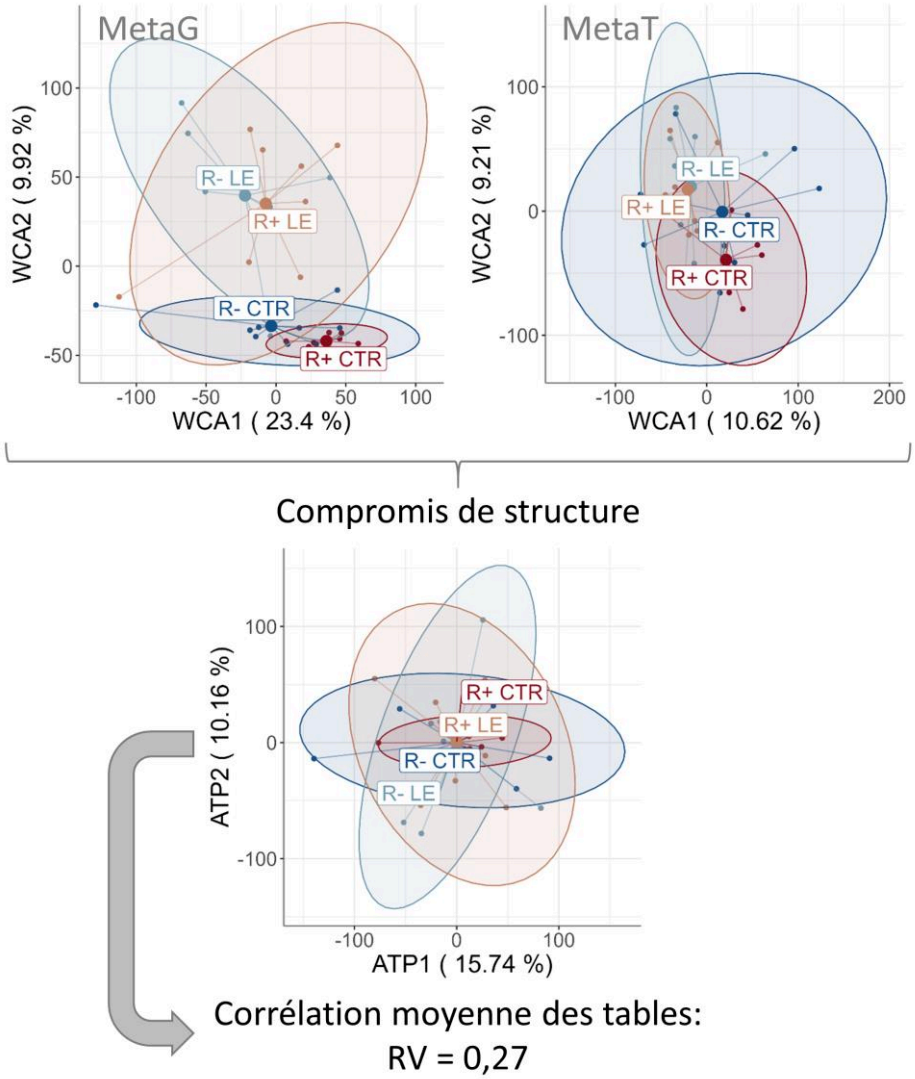


Figure annexe J.2-2 : Analyse triadique partielle sur des analyses intra groupe (WCA). L'analyse WCA est réalisé intra groupe « lignée x régime ».

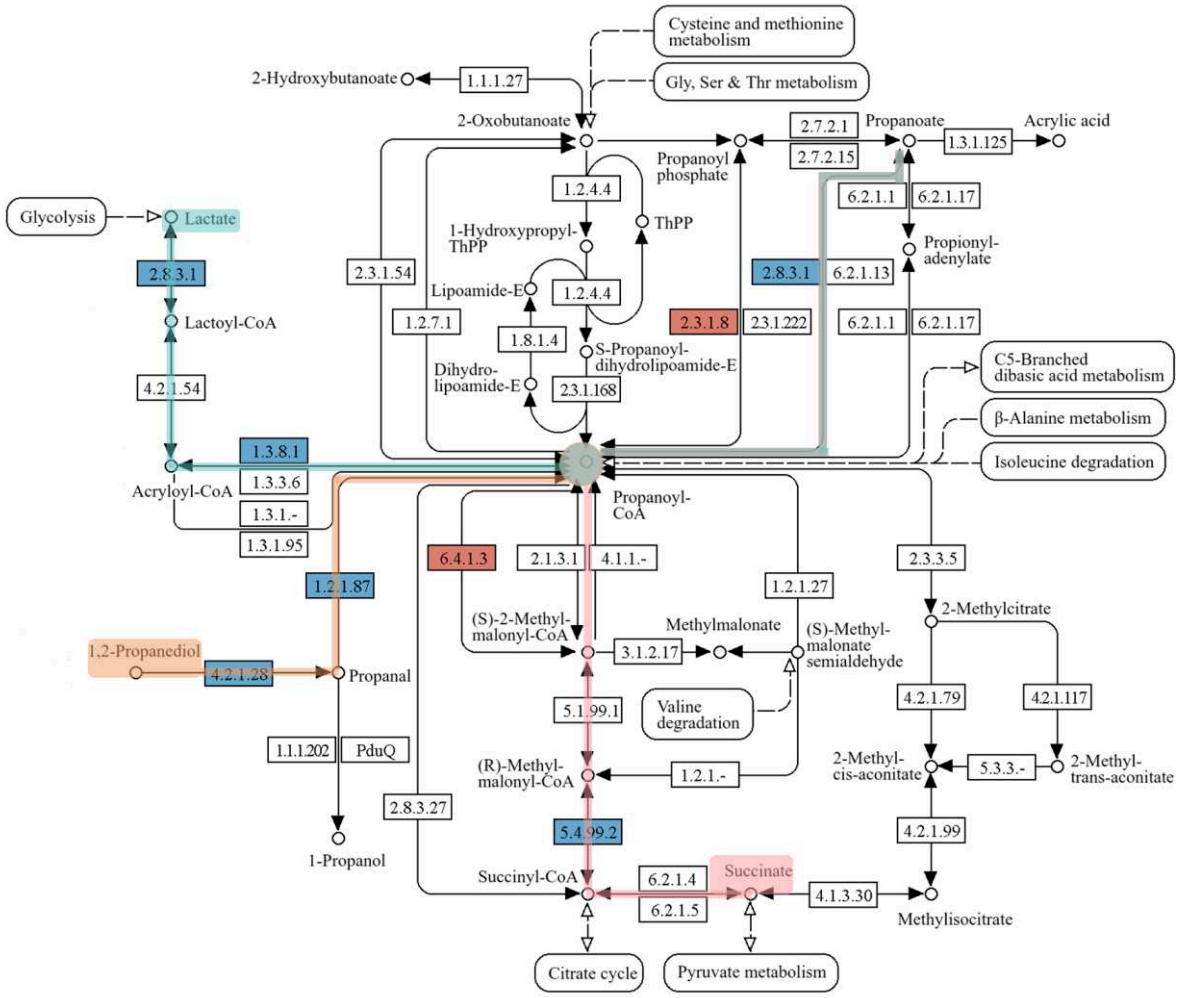


Figure annexe J.2-3: Fonctions du métabolisme du propionate différentiellement abondantes entre lignée (métabarcoding 16S).
Les fonctions et leur abondance sont inférées à partir des séquences la région V3-V4 de l'ARNr 16S. Les abondances sont comparées selon la lignée lorsque les poules sont nourries avec le régime CTR. Les fonctions en bleu sont détectées en surabondance dans la lignée efficiente R-, les fonctions en rouge sont détectées en surabondance dans la lignée non efficiente R+. Résultats issus de la partie 1 des résultats (chapitre **D**). La carte du métabolisme est adaptée de celle produite par KEGG (Kanehisa et al. 2016).

BIBLIOGRAPHIE

Abarenkov, Kessy, R Henrik Nilsson, Karl-Henrik Larsson, Andy F S Taylor, Tom W May, Tobias Guldberg Frøslev, Julia Pawlowska, et al. 2024. 'The UNITE Database for Molecular Identification and Taxonomic Communication of Fungi and Other Eukaryotes: Sequences, Taxa and Classifications Reconsidered'. *Nucleic Acids Research* 52 (D1): D791–97. https://doi.org/10.1093/nar/gkad1039.

Akinyemi, Fisayo, and Deborah Adewole. 2021. 'Environmental Stress in Chickens and the Potential Effectiveness of Dietary Vitamin Supplementation'. *Frontiers in Animal Science* 2 (November). https://doi.org/10.3389/fanim.2021.775311.

Alagawany, M., Sh. S. Elnesr, and M. R. Farag. 2018. 'The Role of Exogenous Enzymes in Promoting Growth and Improving Nutrient Digestibility in Poultry'. *Iranian Journal of Veterinary Research* 19 (3): 157–64.

Albillos, Agustín, Andrea de Gottardi, and María Rescigno. 2020. 'The Gut-Liver Axis in Liver Disease: Pathophysiological Basis for Therapy'. *Journal of Hepatology* 72 (3): 558–77. https://doi.org/10.1016/j.jhep.2019.10.003.

Alexander, Harriet, Sarah K. Hu, Arianna I. Krinos, Maria Pachiadaki, Benjamin J. Tully, Christopher J. Neely, and Taylor Reiter. 2023. 'Eukaryotic Genomes from a Global Metagenomic Data Set Illuminate Trophic Modes and Biogeography of Ocean Plankton'. *mBio* 14 (6): e01676-23. https://doi.org/10.1128/mbio.01676-23.

Almeida, Alexandre, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S. Pollard, et al. 2021. 'A Unified Catalog of 204,938 Reference Genomes from the Human Gut Microbiome'. *Nature Biotechnology* 39 (1): 105–14. https://doi.org/10.1038/s41587-020-0603-3.

Amir, Amnon, Daniel McDonald, Jose A. Navas-Molina, Evguenia Kopylova, James T. Morton, Zhenjiang Zech Xu, Eric P. Kightley, et al. 2017. 'Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns'. *mSystems* 2 (2): e00191-16. https://doi.org/10.1128/mSystems.00191-16.

Anderson, K.E., G.B. Havenstein, P.K. Jenkins, and J. Osborne. 2013. 'Changes in Commercial Laying Stock Performance, 1958-2011: Thirty-Seven Flocks of the North Carolina Random Sample and Subsequent Layer Performance and Management Tests'. *World's Poultry Science Journal* 69 (3): 489–514. https://doi.org/10.1017/S0043933913000536.

Aramaki, Takuya, Romain Blanc-Mathieu, Hisashi Endo, Koichi Ohkubo, Minoru Kanehisa, Susumu Goto, and Hiroyuki Ogata. 2020. 'KofamKOALA: KEGG Ortholog Assignment Based on Profile HMM and Adaptive Score Threshold'. *Bioinformatics* 36 (7): 2251–52. https://doi.org/10.1093/bioinformatics/btz859.

Argüello, Héctor, Jordi Estellé, Finola C. Leonard, Fiona Crispie, Paul D. Cotter, Orla O'Sullivan, Helen Lynch, et al. 2019. 'Influence of the Intestinal Microbiota on Colonization Resistance to Salmonella and the Shedding Pattern of Naturally Exposed Pigs'. *mSystems* 4 (2): 10.1128/msystems.00021-19. https://doi.org/10.1128/msystems.00021-19.

Aruwa, Christiana Eleojo, Charlene Pillay, Martin M. Nyaga, and Saheed Sabiu. 2021. 'Poultry Gut Health – Microbiome Functions, Environmental Impacts, Microbiome Engineering and Advancements in Characterization Technologies'. *Journal of Animal Science and Biotechnology* 12 (December):119. https://doi.org/10.1186/s40104-021-00640-9.

Aryee, Godson, Sarah M. Luecke, Carl R. Dahlen, Kendall C. Swanson, and Samat Amat. 2023. 'Holistic View and Novel Perspective on Ruminal and Extra-Gastrointestinal Methanogens in Cattle'. *Microorganisms* 11 (11): 2746. https://doi.org/10.3390/microorganisms11112746.

Awad, Wageha A., Claudia Hess, and Michael Hess. 2018. 'Re-Thinking the Chicken–Campylobacter Jejuni Interaction: A Review'. *Avian Pathology* 47 (4): 352–63. https://doi.org/10.1080/03079457.2018.1475724.

Bai, Yu, Xingjian Zhou, Jinbiao Zhao, Zhenyu Wang, Hao Ye, Yu Pi, Dongsheng Che, Dandan Han, Shuai Zhang, and Junjun Wang. 2022. 'Sources of Dietary Fiber Affect the SCFA Production and Absorption in the Hindgut of Growing Pigs'. *Frontiers in Nutrition* 8 (January). https://doi.org/10.3389/fnut.2021.719935.

Bain, M. M., Y. Nys, and I.C. Dunn. 2016. 'Increasing Persistency in Lay and Stabilising Egg Quality in Longer Laying Cycles. What Are the Challenges?' *British Poultry Science* 57 (3): 330–38. https://doi.org/10.1080/00071668.2016.1161727.

Bashiardes, Stavros, Gili Zilberman-Schapira, and Eran Elinav. 2016. 'Use of Metatranscriptomics in Microbiome Research'. *Bioinformatics and Biology Insights* 10 (April):19–25. https://doi.org/10.4137/BBI.S34610.

Bedu-Ferrari, Cassandre, Paul Biscarrat, Philippe Langella, and Claire Cherbuy. 2022. 'Prebiotics and the Human Gut Microbiota: From Breakdown Mechanisms to the Impact on Metabolic Health'. *Nutrients* 14 (10): 2096. https://doi.org/10.3390/nu14102096.

Beghini, Francesco, Lauren J McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, et al. 2021. 'Integrating Taxonomic, Functional, and Strain-Level Profiling of Diverse Microbial Communities with bioBakery 3'. Edited by Peter Turnbaugh, Eduardo Franco, and C Titus Brown. *eLife* 10 (May):e65088. https://doi.org/10.7554/eLife.65088.

Benjamini, Yoav, and Yosef Hochberg. 1995. 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing'. *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Berg, Gabriele, Daria Rybakova, Doreen Fischer, Tomislav Cernava, Marie-Christine Champomier Vergès, Trevor Charles, Xiaoyulong Chen, et al. 2020. 'Microbiome Definition Re-Visited: Old Concepts and New Challenges'. *Microbiome* 8 (1): 103. https://doi.org/10.1186/s40168-020-00875-0.

Bernard, Maria, Jean-Luc Coville, Nicolas Bruneau, Deborah Jardet, Fanny Calenge, Géraldine Pascal, and Tatiana Zerjal. 2022. 'PERFORM PROTOCOL TESTS AND SEQUENCE CHECKS ? THE EXPERIENCE FROM MULTIOMICS MICROBIOTA DATA'. September. https://hal.science/hal-03930031.

Bernard, Maria, Alexandre Lecoeur, Jean-Luc Coville, Nicolas Bruneau, Deborah Jardet, Sandrine Lagarrigue, Annabelle Meynadier, Fanny Calenge, Géraldine Pascal, and Tatiana Zerjal. 2024. 'Relationship between Feed Efficiency and Gut Microbiota in Laying Chickens under Contrasting Feeding Conditions'. *Scientific Reports* 14 (1): 8210. https://doi.org/10.1038/s41598-024-58374-3.

Bernard, Maria, Olivier Rué, Mahendra Mariadassou, and Géraldine Pascal. 2021. 'FROGS: A Powerful Tool to Analyse the Diversity of Fungi with Special Management of Internal Transcribed Spacers'. *Briefings in Bioinformatics* 22 (6): bbab318. https://doi.org/10.1093/bib/bbab318.

Berney, Cédric, Andreea Ciuprina, Sara Bender, Juliet Brodie, Virginia Edgcomb, Eunsoo Kim, Jeena Rajan, et al. 2017. 'UniEuk: Time to Speak a Common Language in Protistology!' *Journal of Eukaryotic Microbiology* 64 (3): 407–11. https://doi.org/10.1111/jeu.12414.

Bharti, Richa, and Dominik G Grimm. 2021. 'Current Challenges and Best-Practice Protocols for Microbiome Analysis'. *Briefings in Bioinformatics* 22 (1): 178–93. https://doi.org/10.1093/bib/bbz155.

Bindari, Yugal Raj, and Priscilla F. Gerber. 2022. 'Centennial Review: Factors Affecting the Chicken Gastrointestinal Microbial Composition and Their Association with Gut Health and Productive Performance'. *Poultry Science* 101 (1): 101612. https://doi.org/10.1016/j.psj.2021.101612.

Birhanu, Mulugeta Y, Richard Osei-Amponsah, Frederick Yeboah Obese, and Tadelle Dessie. 2023. 'Smallholder Poultry Production in the Context of Increasing Global Food Prices: Roles in

Poverty Reduction and Food Security'. *Animal Frontiers* 13 (1): 17–25. https://doi.org/10.1093/af/vfac069.

Birkl, P, A Bharwani, J. B. Kjaer, W Kunze, P McBride, P Forsythe, and A Harlander-Matauschek. 2018. 'Differences in Cecal Microbiome of Selected High and Low Feather-Pecking Laying Hens'. *Poultry Science* 97 (9): 3009–14. https://doi.org/10.3382/ps/pey167.

Blanco-Miguez, Aitor, Francesco Beghini, Fabio Cumbo, Lauren J. McIver, Kelsey N. Thompson, Moreno Zolfo, Paolo Manghi, et al. 2022. 'Extending and Improving Metagenomic Taxonomic Profiling with Uncharacterized Species with MetaPhlAn 4'. bioRxiv. https://doi.org/10.1101/2022.08.22.504593.

Bokulich, Nicholas A, Sathish Subramanian, Jeremiah J Faith, Dirk Gevers, Jeffrey I Gordon, Rob Knight, David A Mills, and J Gregory Caporaso. 2013. 'Quality-Filtering Vastly Improves Diversity Estimates from Illumina Amplicon Sequencing'. *Nature Methods* 10 (1): 57–59. https://doi.org/10.1038/nmeth.2276.

Bonk, Fabian, Denny Popp, Hauke Harms, and Florian Centler. 2018. 'PCR-Based Quantification of Taxa-Specific Abundances in Microbial Communities: Quantifying and Avoiding Common Pitfalls'. *Journal of Microbiological Methods* 153 (October):139–47. https://doi.org/10.1016/j.mimet.2018.09.015.

Bordas, A., and P. Mérat. 1974. 'Variabilité Génétique et Corrélations Phénotypiques Caractérisant La Consommation Alimentaire de Poules Pondeuses Après Correction Pour Le Poids Corporel et La Ponte'. *Annales de Génétique et de Sélection Animale* 6 (3): 369. https://doi.org/10.1186/1297-9686-6-3-369.

———. 1984. 'Correlated Responses in a Selection Experiment on Residual Feed Intake of Adult Rhode-Island Red Cocks and Hens'. *Ann. Agric. Fenn* 23:233--237.

Bordas, A., P. Mérat, G. Coquerelle, and JP Noé. 1995. 'Influence d'un Aliment Dilué Sur Des Lignées de Poules Pondeuses Sélectionnées Sur La Consommation Alimentaire Résiduelle'. *Genetics Selection Evolution* 27 (3): 299. https://doi.org/10.1186/1297-9686-27-3-299.

Bordas, A., and F. Minvielle. 1997. 'Réponse à La Chaleur de Poules Pondeuses Issues de Lignées Sélectionnées Pour Une Faible (R-) Ou Forte (R+) Consommation Alimentaire Résiduelle'. *Genetics Selection Evolution* 29 (2): 279. https://doi.org/10.1186/1297-9686-29-2-279.

Bordas, A., M. Tixier-Boichard, and P. Merat. 1992. 'Direct and Correlated Responses to Divergent Selection for Residual Food Intake in Rhode Island Red Laying Hens'. *British Poultry Science* 33 (4): 741–54. https://doi.org/10.1080/00071669208417515.

Bordenstein, Seth R., and Kevin R. Theis. 2015. 'Host Biology in Light of the Microbiome: Ten Principles of Holobionts and Hologenomes'. *PLoS Biology* 13 (8): e1002226. https://doi.org/10.1371/journal.pbio.1002226.

Borey, Marion, Jordi Estellé, Aziza Caidi, Nicolas Bruneau, Jean-Luc Coville, Christelle Hennequet-Antier, Sandrine Mignon-Grasteau, and Fanny Calenge. 2020. 'Broilers Divergently Selected for Digestibility Differ for Their Digestive Microbial Ecosystems'. *PLOS ONE* 15 (5): e0232418. https://doi.org/10.1371/journal.pone.0232418.

Brandt, Miriam I., Blandine Trouche, Laure Quintric, Babett Günther, Patrick Wincker, Julie Poulain, and Sophie Arnaud-Haond. 2021. 'Bioinformatic Pipelines Combining Denoising and Clustering Tools Allow for More Comprehensive Prokaryotic and Eukaryotic Metabarcoding'. *Molecular Ecology Resources* 21 (6): 1904–21. https://doi.org/10.1111/1755-0998.13398.

Breitwieser, Florian P, Jennifer Lu, and Steven L Salzberg. 2019. 'A Review of Methods and Databases for Metagenomic Classification and Assembly'. *Briefings in Bioinformatics* 20 (4): 1125–36. https://doi.org/10.1093/bib/bbx120.

Brumfield, Kyle D., Anwar Huq, Rita R. Colwell, James L. Olds, and Menu B. Leddy. 2020. 'Microbial Resolution of Whole Genome Shotgun and 16S Amplicon Metagenomic Sequencing Using Publicly Available NEON Data'. *PLOS ONE* 15 (2): e0228899. https://doi.org/10.1371/journal.pone.0228899.

Bubier, Jason A., Elissa J. Chesler, and George M. Weinstock. 2021. 'Host Genetic Control of Gut Microbiome Composition'. *Mammalian Genome* 32 (4): 263–81. https://doi.org/10.1007/s00335-021-09884-2.

Buchfink, Benjamin, Klaus Reuter, and Hajk-Georg Drost. 2021. 'Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND'. *Nature Methods* 18 (4): 366–68. https://doi.org/10.1038/s41592-021-01101-x.

Buetas, Elena, Marta Jordán-López, Andrés López-Roldán, Giuseppe D'Auria, Llucia Martínez-Priego, Griselda De Marco, Miguel Carda-Diéguez, and Alex Mira. 2024. 'Full-Length 16S rRNA Gene Sequencing by PacBio Improves Taxonomic Resolution in Human Microbiome Samples'. *BMC Genomics* 25 (1): 310. https://doi.org/10.1186/s12864-024-10213-5.

Byerly, Theodore Carroll. 1941. *Feed and Other Costs of Producing Market Eggs*. University of Maryland, Agricultural Experiment Station.

Cahana, Inbal, and Fuad A. Iraqi. 2020. 'Impact of Host Genetics on Gut Microbiome: Take-home Lessons from Human and Mouse Studies'. *Animal Models and Experimental Medicine* 3 (3): 229–36. https://doi.org/10.1002/ame2.12134.

Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. 2016. 'DADA2: High Resolution Sample Inference from Illumina Amplicon Data'. *Nature Methods* 13 (7): 581–83. https://doi.org/10.1038/nmeth.3869.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. 2009. 'BLAST+: Architecture and Applications'. *BMC Bioinformatics* 10 (1): 421. https://doi.org/10.1186/1471-2105-10-421.

Campbell, Anthony K., Jonathan P. Waud, and Stephanie B. Matthews. 2009. 'The Molecular Basis of Lactose Intolerance'. *Science Progress* 92 (Pt 3-4): 241–87. https://doi.org/10.3184/003685009X12547510332240.

Cantalapiedra-Hijar, Gonzalo, Philippe Faverdin, Nicolas C. Friggens, and Pauline Martin. 2020. 'Efficience Alimentaire : comment mieux la comprendre et en faire un élément de durabilité de l'élevage'. *INRAE Productions Animales* 33 (4): 235–48. https://doi.org/10.20870/productions-animales.2020.33.4.4594.

Cantu-Jungles, Thaisa M., and Bruce R. Hamaker. 2023. 'Tuning Expectations to Reality: Don't Expect Increased Gut Microbiota Diversity with Dietary Fiber'. *The Journal of Nutrition* 153 (11): 3156–63. https://doi.org/10.1016/j.tjnut.2023.09.001.

Carini, Paul, Patrick J. Marsden, Jonathan W. Leff, Emily E. Morgan, Michael S. Strickland, and Noah Fierer. 2016. 'Relic DNA Is Abundant in Soil and Obscures Estimates of Soil Microbial Diversity'. *Nature Microbiology* 2 (3): 1–6. https://doi.org/10.1038/nmicrobiol.2016.242.

Carmona-Cruz, Silvia, Luz Orozco-Covarrubias, and Marimar Sáez-de-Ocariz. 2022. 'The Human Skin Microbiome in Selected Cutaneous Diseases'. *Frontiers in Cellular and Infection Microbiology* 12 (March). https://doi.org/10.3389/fcimb.2022.834135.

Casey, Jordan M., Emma Ransome, Allen G. Collins, Angka Mahardini, Eka M. Kurniasih, Andrianus Sembiring, Nina M. D. Schiettekatte, et al. 2021. 'DNA Metabarcoding Marker Choice Skews Perception of Marine Eukaryotic Biodiversity'. *Environmental DNA* 3 (6): 1229–46. https://doi.org/10.1002/edn3.245.

Chaumeil, Pierre-Alain, Aaron J Mussig, Philip Hugenholtz, and Donovan H Parks. 2020. 'GTDB-Tk: A Toolkit to Classify Genomes with the Genome Taxonomy Database'. *Bioinformatics* 36 (6): 1925–27. https://doi.org/10.1093/bioinformatics/btz848.

Chen, Congying, Yunyan Zhou, Hao Fu, Xinwei Xiong, Shaoming Fang, Hui Jiang, Jinyuan Wu, Hui Yang, Jun Gao, and Lusheng Huang. 2021. 'Expanded Catalog of Microbial Genes and Metagenome-Assembled Genomes from the Pig Gut Microbiome'. *Nature Communications* 12 (1): 1106. https://doi.org/10.1038/s41467-021-21295-0.

Chklovski, Alex, Donovan H. Parks, Ben J. Woodcroft, and Gene W. Tyson. 2022. 'CheckM2: A Rapid, Scalable and Accurate Tool for Assessing Microbial Genome Quality Using Machine Learning'. bioRxiv. https://doi.org/10.1101/2022.07.11.499243.

Cho, Hunyong, Yixiang Qu, Chuwen Liu, Boyang Tang, Ruiqi Lyu, Bridget M Lin, Jeffrey Roach, et al. 2023. 'Comprehensive Evaluation of Methods for Differential Expression Analysis of Metatranscriptomics Data'. *Briefings in Bioinformatics* 24 (5): bbad279. https://doi.org/10.1093/bib/bbad279.

Chungchunlam, Sylvia M. S., and Paul J. Moughan. 2023. 'Comparative Bioavailability of Vitamins in Human Foods Sourced from Animals and Plants'. *Critical Reviews in Food Science and Nutrition*, July, 1–36. https://doi.org/10.1080/10408398.2023.2241541.

Cisek, Agata Anna, Beata Dolka, Iwona Bąk, and Bożena Cukrowska. 2023. 'Microorganisms Involved in Hydrogen Sink in the Gastrointestinal Tract of Chickens'. *International Journal of Molecular Sciences* 24 (7): 6674. https://doi.org/10.3390/ijms24076674.

Clark, Allison, and Núria Mach. 2016. 'Exercise-Induced Stress Behavior, Gut-Microbiota-Brain Axis and Diet: A Systematic Review for Athletes'. *Journal of the International Society of Sports Nutrition* 13 (1): 43. https://doi.org/10.1186/s12970-016-0155-6.

Cole, James R., Qiong Wang, Jordan A. Fish, Benli Chai, Donna M. McGarrell, Yanni Sun, C. Titus Brown, Andrea Porras-Alfaro, Cheryl R. Kuske, and James M. Tiedje. 2014. 'Ribosomal Database Project: Data and Tools for High Throughput rRNA Analysis'. *Nucleic Acids Research* 42 (D1): D633–42. https://doi.org/10.1093/nar/gkt1244.

Da Silva, Luís P., Vanessa A. Mata, Pedro B. Lopes, Paulo Pereira, Simon N. Jarman, Ricardo J. Lopes, and Pedro Beja. 2019. 'Advancing the Integration of Multi-marker Metabarcoding Data in Dietary Analysis of Trophic Generalists'. *Molecular Ecology Resources* 19 (6): 1420–32. https://doi.org/10.1111/1755-0998.13060.

Danzeisen, Jessica L., Hyeun Bum Kim, Richard E. Isaacson, Zheng Jin Tu, and Timothy J. Johnson. 2011. 'Modulations of the Chicken Cecal Microbiome and Metagenome in Response to Anticoccidial and Growth Promoter Treatment'. *PloS One* 6 (11): e27949. https://doi.org/10.1371/journal.pone.0027949.

Darbot, Vincent, Moussa Samb, Maria Bernard, Olivier Rué, and Géraldine Pascal. 2022. 'FROGSFUNC: Smart Integration of PICRUSt2 Software into FROGS Pipeline'. In . https://hal.inrae.fr/hal-03806133.

Dart, R. K. 1996. *Microbiology for the Analytical Chemist*. The Royal Society of Chemistry. https://doi.org/10.1039/9781847551443.

Darwish, Nadia, Jonathan Shao, Lori L. Schreier, and Monika Proszkowiec-Weglarz. 2021. 'Choice of 16S Ribosomal RNA Primers Affects the Microbiome Analysis in Chicken Ceca'. *Scientific Reports* 11 (1): 11848. https://doi.org/10.1038/s41598-021-91387-w.

Day, Li, Julie A. Cakebread, and Simon M. Loveday. 2022. 'Food Proteins from Animals and Plants: Differences in the Nutritional and Functional Properties'. *Trends in Food Science & Technology* 119 (January):428–42. https://doi.org/10.1016/j.tifs.2021.12.020.

De Filippis, Francesca, Manolo Laiola, Giuseppe Blaiotta, and Danilo Ercolini. 2017. 'Different Amplicon Targets for Sequencing-Based Studies of Fungal Diversity'. *Applied and Environmental Microbiology* 83 (17): e00905-17. https://doi.org/10.1128/AEM.00905-17.

Desbruslais, Alexandra, Alexandra Wealleans, David Gonzalez-Sanchez, and Mauro di Benedetto. 2021. 'Dietary Fibre in Laying Hens: A Review of Effects on Performance, Gut Health and Feather Pecking'. *World's Poultry Science Journal* 77 (4): 797–823. https://doi.org/10.1080/00439339.2021.1960236.

Destrez, Alexandra, Pauline Grimm, Frank Cézilly, and Véronique Julliand. 2015. 'Changes of the Hindgut Microbiota Due to High-Starch Diet Can Be Associated with Behavioral Stress Response in Horses'. *Physiology & Behavior* 149 (October):159–64. https://doi.org/10.1016/j.physbeh.2015.05.039.

Deusch, Simon, Bruno Tilocca, Amélia Camarinha-Silva, and Jana Seifert. 2015. 'News in Livestock Research — Use of *Omics*-Technologies to Study the Microbiota in the Gastrointestinal Tract of Farm Animals'. *Computational and Structural Biotechnology Journal* 13 (January):55–63. https://doi.org/10.1016/j.csbj.2014.12.005.

DeVries, T. J., and E. Chevaux. 2014. 'Modification of the Feeding Behavior of Dairy Cows through Live Yeast Supplementation'. *Journal of Dairy Science* 97 (10): 6499–6510. https://doi.org/10.3168/jds.2014-8226.

Diakite, Ami, Grégory Dubourg, and Didier Raoult. 2021. 'Updating the Repertoire of Cultured Bacteria from the Human Being'. *Microbial Pathogenesis* 150 (January):104698. https://doi.org/10.1016/j.micpath.2020.104698.

Dijk, Michiel van, Tom Morley, Marie Luise Rau, and Yashar Saghai. 2021. 'A Meta-Analysis of Projected Global Food Demand and Population at Risk of Hunger for the Period 2010–2050'. *Nature Food* 2 (7): 494–501. https://doi.org/10.1038/s43016-021-00322-9.

Dinan, Timothy G., and John F. Cryan. 2017. 'Brain–Gut–Microbiota Axis — Mood, Metabolism and Behaviour'. *Nature Reviews Gastroenterology & Hepatology* 14 (2): 69–70. https://doi.org/10.1038/nrgastro.2016.200.

Donkey shot. 2016. *Gallus Sonneratii*. Map Source www.shadedrelief.com Data Sources J. del Hoyo, A. Elliott, J. Sargatal, A. D. Christie, E. de Juana (Hg.): Handbook of the Birds of the World Alive. Lynx Edicions, Barcelona 2016 The IUCN Red List of Threatened Species, 2015-4. https://commons.wikimedia.org/wiki/File:Gallus_distribution.jpg.

Douglas, Gavin M., and Morgan G. I. Langille. 2021. 'A primer and discussion on DNA-based microbiome data and related bioinformatics analyses'. *Peer Community Journal* 1. https://doi.org/10.24072/pcjournal.2.

Douglas, Gavin M., Vincent J. Maffei, Jesse R. Zaneveld, Svetlana N. Yurgel, James R. Brown, Christopher M. Taylor, Curtis Huttenhower, and Morgan G. I. Langille. 2020. 'PICRUSt2 for Prediction of Metagenome Functions'. *Nature Biotechnology* 38 (6): 685–88. https://doi.org/10.1038/s41587-020-0548-6.

Dray, Stéphane, Anne B Dufour, and Daniel Chessel. 2007. 'The Ade4 Package — II: Two-Table and K-Table Methods' 7.

Drula, Elodie, Marie-Line Garron, Suzan Dogan, Vincent Lombard, Bernard Henrissat, and Nicolas Terrapon. 2022. 'The Carbohydrate-Active Enzyme Database: Functions and Literature'. *Nucleic Acids Research* 50 (D1): D571–77. https://doi.org/10.1093/nar/gkab1045.

Dueholm, Morten Kam Dahl, Kasper Skytte Andersen, Anne-Kirstine C. Petersen, Vibeke Rudkjøbing, Madalena Alves, Yadira Bajón-Fernández, Damien Batstone, et al. 2024. 'MiDAS 5: Global Diversity of Bacteria and Archaea in Anaerobic Digesters'. bioRxiv. https://doi.org/10.1101/2023.08.24.554448.

Durazzi, Francesco, Claudia Sala, Gastone Castellani, Gerardo Manfreda, Daniel Remondini, and Alessandra De Cesare. 2021. 'Comparison between 16S rRNA and Shotgun Sequencing Data for the Taxonomic Characterization of the Gut Microbiota'. *Scientific Reports* 11 (1): 3030. https://doi.org/10.1038/s41598-021-82726-y.

Ebertz, Dr Andreas. 2020. 'A Journey Through The History Of DNA Sequencing'. *The DNA Universe BLOG* (blog). 2 November 2020. https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/.

Edgar, Robert C. 2013. 'UPARSE: Highly Accurate OTU Sequences from Microbial Amplicon Reads'. *Nature Methods* 10 (10): 996–98. https://doi.org/10.1038/nmeth.2604.

Edgar, Robert C. 2018. 'Updating the 97% Identity Threshold for 16S Ribosomal RNA OTUs'. *Bioinformatics* 34 (14): 2371–75. https://doi.org/10.1093/bioinformatics/bty113.

Eisenhofer, Raphael, Joseph Nesme, Luisa Santos-Bay, Adam Koziol, Søren Johannes Sørensen, Antton Alberdi, and Ostaizka Aizpurua. 2024. 'A Comparison of Short-Read, HiFi Long-Read, and Hybrid Strategies for Genome-Resolved Metagenomics'. *Microbiology Spectrum* 12 (4): e03590-23. https://doi.org/10.1128/spectrum.03590-23.

Elem, Oghenekaro. 2024. 'Feed Conversion Ratio Calculator | FCR'. Omni Calculator. 26 April 2024. https://www.omnicalculator.com/biology/fcr.

El-Kazzi, M., A. Bordas, G. Gandemer, and F. Minvielle. 1995a. 'Divergent Selection for Residual Food Intake in Rhode Island Red Egg-Laying Lines: Gross Carcase Composition, Carcase Adiposity and

Lipid Contents of Tissues'. *British Poultry Science* 36 (5): 719–28. https://doi.org/10.1080/00071669508417816.

———. 1995b. 'Divergent Selection for Residual Food Intake in Rhode Island Red Egg-Laying Lines: Gross Carcase Composition, Carcase Adiposity and Lipid Contents of Tissues'. *British Poultry Science* 36 (5): 719–28. https://doi.org/10.1080/00071669508417816.

Escudié, Frédéric, Lucas Auer, Maria Bernard, Mahendra Mariadassou, Laurent Cauquil, Katia Vidal, Sarah Maman, Guillermina Hernandez-Raquet, Sylvie Combes, and Géraldine Pascal. 2018. 'FROGS: Find, Rapidly, OTUs with Galaxy Solution'. *Bioinformatics (Oxford, England)* 34 (8): 1287–94. https://doi.org/10.1093/bioinformatics/btx791.

Fan, Peixin, Beilei Bian, Lin Teng, Corwin D. Nelson, J. Driver, Mauricio A. Elzo, and Kwangcheol C. Jeong. 2020. 'Host Genetic Effects upon the Early Gut Microbiota in a Bovine Model with Graduated Spectrum of Genetic Variation'. *The ISME Journal* 14 (1): 302–17. https://doi.org/10.1038/s41396-019-0529-2.

Fang, Shaoming, Xuan Chen, Xiaoxing Ye, Liwen Zhou, Shuaishuai Xue, and Qianfu Gan. 2020. 'Effects of Gut Microbiome and Short-Chain Fatty Acids (SCFAs) on Finishing Weight of Meat Rabbits'. *Frontiers in Microbiology* 11 (August):1835. https://doi.org/10.3389/fmicb.2020.01835.

FAO. 2023a. *Contribution of Terrestrial Animal Source Food to Healthy Diets for Improved Nutrition and Health Outcomes*. FAO. https://doi.org/10.4060/cc3912en.

———. 2023b. 'Food and Agriculture Organization of the United Nations - Production: Crops and Livestock Products'. http://www.fao.org/faostat/en/#data/QCL. http://www.fao.org/faostat/en/#data/QCL.

Feng, Yuqing, Yanan Wang, Baoli Zhu, George Fu Gao, Yuming Guo, and Yongfei Hu. 2021. 'Metagenome-Assembled Genomes and Gene Catalog from the Chicken Gut Microbiome Aid in Deciphering Antibiotic Resistomes'. *Communications Biology* 4 (1): 1–9. https://doi.org/10.1038/s42003-021-02827-2.

Fondation Tara Océan. 2020. 'Comprendre le peuple invisible de l'Océan | Mission Microbiomes'. Fondation Tara Océan. 2020. https://fondationtaraocean.org/expedition/mission-microbiomes/.

Fouhse, J.M., R.T. Zijlstra, and B.P. Willing. 2016. 'The Role of Gut Microbiota in the Health and Disease of Pigs'. *Animal Frontiers* 6 (3): 30–36. https://doi.org/10.2527/af.2016-0031.

Fourquet, Joanna, Jean Mainguy, Maïna Vienne, Céline Noirot, Pierre Martin, Vincent Darbot, Olivier Bouchez, et al. 2022. 'metagWGS: A Workflow to Analyse Short and Long HiFi Metagenomic Reads Taxonomic Profile HiFi vs Short Reads Assembly'. In . https://doi.org/10.15454/1.5572369328961167E12.

Fox, John, Sanford Weisberg, Brad Price, Daniel Adler, Douglas Bates, Gabriel Baud-Bovy, Ben Bolker, et al. 2022. 'Car: Companion to Applied Regression'. https://CRAN.R-project.org/package=car.

Franzosa, Eric A., Xochitl C. Morgan, Nicola Segata, Levi Waldron, Joshua Reyes, Ashlee M. Earl, Georgia Giannoukos, et al. 2014. 'Relating the Metatranscriptome and Metagenome of the Human Gut'. *Proceedings of the National Academy of Sciences of the United States of America* 111 (22): E2329–38. https://doi.org/10.1073/pnas.1319284111.

Gabarrou, Jean-François, P.A. Geraert, N. Francois, S. Guillaumin, M. Picard, and A. Bordas. 1998. 'Energy Balance of Laying Hens Selected on Residual Food Consumption'. *British Poultry Science* 39 (1): 79–89. https://doi.org/10.1080/00071669889439.

Gabarrou, Jean-François, Pierre-André Géraert, Michel Picard, and André Bordas. 1997. 'Diet-Induced Thermogenesis in Cockerels Is Modulated by Genetic Selection for High or Low Residual Feed Intake'. *The Journal of Nutrition* 127 (12): 2371–76. https://doi.org/10.1093/jn/127.12.2371.

Gaci, Nadia, Guillaume Borrel, William Tottey, Paul William O'Toole, and Jean-François Brugère. 2014. 'Archaea and the Human Gut: New Beginning of an Old Story'. *World Journal of Gastroenterology : WJG* 20 (43): 16062–78. https://doi.org/10.3748/wjg.v20.i43.16062.

Gardiner, Gillian E., Barbara U. Metzler-Zebeli, and Peadar G. Lawlor. 2020. 'Impact of Intestinal Microbiota on Growth and Feed Efficiency in Pigs: A Review'. *Microorganisms* 8 (12): 1886. https://doi.org/10.3390/microorganisms8121886.

Garrity, George M, Julia A Bell, and Timothy G Lilburn. 2004. 'TAXONOMIC OUTLINE OF THE PROKARYOTES BERGEY'S MANUAL® OF SYSTEMATIC BACTERIOLOGY, SECOND EDITION', May.

Gaudichon, Claire, and Juliane Calvez. 2021. 'Determinants of Amino Acid Bioavailability from Ingested Protein in Relation to Gut Health'. *Current Opinion in Clinical Nutrition and Metabolic Care* 24 (1): 55–61. https://doi.org/10.1097/MCO.0000000000000708.

GBIF Secretariat. 2023. 'GBIF Backbone Taxonomy'. https://doi.org/10.15468/39omei.

George, E. Olusegun, and Govind S. Mudholkar. 1990. 'P-Values for Two-Sided Tests'. *Biometrical Journal* 32 (6): 747–51. https://doi.org/10.1002/bimj.4710320615.

Gevers, Dirk, Mihai Pop, Patrick D. Schloss, and Curtis Huttenhower. 2012. 'Bioinformatics for the Human Microbiome Project'. *PLOS Computational Biology* 8 (11): e1002779. https://doi.org/10.1371/journal.pcbi.1002779.

Gilbert, H, J P Bidanel, J Gruand, J C Caritez, Y Billon, P Guillouet, J Noblet, and P Sellier. 2006. 'GENETIC PARAMETERS AND RESPONSES TO DIVERGENT SELECTION FOR RESIDUAL FEED INTAKE IN THE GROWING PIG'. In . Belo Horizonte, MG, Brasil.

Gilbert, Jack A., Janet K. Jansson, and Rob Knight. 2014. 'The Earth Microbiome Project: Successes and Aspirations'. *BMC Biology* 12 (1): 69. https://doi.org/10.1186/s12915-014-0069-1.

Gilbert, Marius, Giuseppina Cinardi, Daniele Da Re, William G. R. Wint, Dominik Wisser, and Timothy P. Robinson. 2022. 'Global Chickens Distribution in 2015 (5 Minutes of Arc)'. Harvard Dataverse. https://doi.org/10.7910/DVN/SXHLF3.

Gilbert, Marius, Gaëlle Nicolas, Giusepina Cinardi, Thomas P. Van Boeckel, Sophie O. Vanwambeke, G. R. William Wint, and Timothy P. Robinson. 2018. 'Global Distribution Data for Cattle, Buffaloes, Horses, Sheep, Goats, Pigs, Chickens and Ducks in 2010'. *Scientific Data* 5 (1): 180227. https://doi.org/10.1038/sdata.2018.227.

Gilroy, Rachel, Anuradha Ravi, Maria Getino, Isabella Pursley, Daniel L. Horton, Nabil-Fareed Alikhan, Dave Baker, et al. 2021. 'Extensive Microbial Diversity within the Chicken Gut Microbiome Revealed by Metagenomics and Culture'. *PeerJ* 9 (April):e10941. https://doi.org/10.7717/peerj.10941.

Girardie, Océane, Denis Laloë, Mathieu Bonneau, Yvon Billon, Jean Bailly, Ingrid David, and Laurianne Canario. 2024. 'Primiparous Sow Behaviour on the Day of Farrowing as One of the Primary Contributors to the Growth of Piglets in Early Lactation'. *Scientific Reports* 14 (1): 18415. https://doi.org/10.1038/s41598-024-69358-8.

Glendinning, Laura, Robert D. Stewart, Mark J. Pallen, Kellie A. Watson, and Mick Watson. 2020. 'Assembly of Hundreds of Novel Bacterial Genomes from the Chicken Caecum'. *Genome Biology* 21 (1): 34. https://doi.org/10.1186/s13059-020-1947-1.

Goodrich, Julia K., Emily R. Davenport, Jillian L. Waters, Andrew G. Clark, and Ruth E. Ley. 2016. 'Cross-Species Comparisons of Host Genetic Associations with the Microbiome'. *Science (New York, N.Y.)* 352 (6285): 532–35. https://doi.org/10.1126/science.aad9379.

Gosalbes, María José, Ana Durbán, Miguel Pignatelli, Juan José Abellan, Nuria Jiménez-Hernández, Ana Elena Pérez-Cobas, Amparo Latorre, and Andrés Moya. 2011. 'Metatranscriptomic Approach to Analyze the Functional Human Gut Microbiota'. *PloS One* 6 (3): e17447. https://doi.org/10.1371/journal.pone.0017447.

Grau-Del Valle, Carmen, Javier Fernández, Eva Solá, Inmaculada Montoya-Castilla, Carlos Morillas, and Celia Bañuls. 2023. 'Association between Gut Microbiota and Psychiatric Disorders: A Systematic Review'. *Frontiers in Psychology* 14 (August). https://doi.org/10.3389/fpsyg.2023.1215674.

Groussin, Mathieu, Florent Mazel, and Eric J. Alm. 2020. 'Co-Evolution and Co-Speciation of Host-Gut Bacteria Systems'. *Cell Host & Microbe* 28 (1): 12–22. https://doi.org/10.1016/j.chom.2020.06.013.

Gulyás, Gábor, Balázs Kakuk, Ákos Dörmő, Tamás Járay, István Prazsák, Zsolt Csabai, Miksa Máté Henkrich, Zsolt Boldogkői, and Dóra Tombácz. 2023. 'Cross-Comparison of Gut Metagenomic Profiling Strategies'. bioRxiv. https://doi.org/10.1101/2023.11.25.568646.

Hakimzadeh, Ali, Alejandro Abdala Asbun, Davide Albanese, Maria Bernard, Dominik Buchner, Benjamin Callahan, J. Gregory Caporaso, et al. 2023. 'A Pile of Pipelines: An Overview of the Bioinformatics Software for Metabarcoding Data Analyses'. *Molecular Ecology Resources*, August, 1755-0998.13847. https://doi.org/10.1111/1755-0998.13847.

Hassler, Hayley B., Brett Probert, Carson Moore, Elizabeth Lawson, Richard W. Jackson, Brook T. Russell, and Vincent P. Richards. 2022. 'Phylogenies of the 16S rRNA Gene and Its Hypervariable Regions Lack Concordance with Core Genome Phylogenies'. *Microbiome* 10 (July):104. https://doi.org/10.1186/s40168-022-01295-y.

Havenstein, G. B., P. R. Ferket, and M. A. Qureshi. 2003. 'Growth, Livability, and Feed Conversion of 1957 versus 2001 Broilers When Fed Representative 1957 and 2001 Broiler Diets'. *Poultry Science* 82 (10): 1500–1508. https://doi.org/10.1093/ps/82.10.1500.

He, Zhengxiao, Ranran Liu, Mengjie Wang, Qiao Wang, Jumei Zheng, Jiqiang Ding, Jie Wen, Alan G. Fahey, and Guiping Zhao. 2023. 'Combined Effect of Microbially Derived Cecal SCFA and Host Genetics on Feed Efficiency in Broiler Chickens'. *Microbiome* 11 (1): 198. https://doi.org/10.1186/s40168-023-01627-6.

Heintz-Buschart, Anna, Patrick May, Cédric C. Laczny, Laura A. Lebrun, Camille Bellora, Abhimanyu Krishna, Linda Wampach, et al. 2016. 'Integrated Multi-Omics of the Human Gut Microbiome in a Case Study of Familial Type 1 Diabetes'. *Nature Microbiology* 2 (October):16180. https://doi.org/10.1038/nmicrobiol.2016.180.

Heintz-Buschart, Anna, and Paul Wilmes. 2018. 'Human Gut Microbiome: Function Matters'. *Trends in Microbiology* 26 (7): 563–74. https://doi.org/10.1016/j.tim.2017.11.002.

Hooks, Katarzyna B., and Maureen A. O'Malley. 2020. 'Contrasting Strategies: Human Eukaryotic Versus Bacterial Microbiome Research'. *Journal of Eukaryotic Microbiology* 67 (2): 279–95. https://doi.org/10.1111/jeu.12766.

Horvathova, Kristyna, Nikol Modrackova, Igor Splichal, Alla Splichalova, Ahmad Amin, Eugenio Ingribelli, Jiri Killer, et al. 2023. 'Defined Pig Microbiota with a Potential Protective Effect against Infection with Salmonella Typhimurium'. *Microorganisms* 11 (4): 1007. https://doi.org/10.3390/microorganisms11041007.

Huang, Peng, Yan Zhang, Kangpeng Xiao, Fan Jiang, Hengchao Wang, Dazhi Tang, Dan Liu, et al. 2018. 'The Chicken Gut Metagenome and the Modulatory Effects of Plant-Derived Benzylisoquinoline Alkaloids'. *Microbiome* 6 (1): 211. https://doi.org/10.1186/s40168-018-0590-5.

Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, et al. 2019. 'eggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses'. *Nucleic Acids Research* 47 (D1): D309–14. https://doi.org/10.1093/nar/gky1085.

Human Microbiome Project Consortium. 2012. 'A Framework for Human Microbiome Research'. *Nature* 486 (7402): 215–21. https://doi.org/10.1038/nature11209.

———. 2014. 'The Integrative Human Microbiome Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of Human Health and Disease'. *Cell Host & Microbe* 16 (3): 276–89. https://doi.org/10.1016/j.chom.2014.08.014.

Hussein, Marwa A., Farina Khattak, Lonneke Vervelde, Spiridoula Athanasiadou, and Jos G. M. Houdijk. 2023. 'Growth Performance, Caecal Microbiome Profile, Short-Chain Fatty Acids, and Litter Characteristics in Response to Placement on Reused Litter and Combined Threonine, Arginine and Glutamine Supplementation to Juvenile Male Broiler Chickens'. *Animal Microbiome* 5 (1): 18. https://doi.org/10.1186/s42523-023-00240-0.

Hyatt, Doug, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. 'Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification'. *BMC Bioinformatics* 11 (1): 119. https://doi.org/10.1186/1471-2105-11-119.

Hy-line. 2024. 'Hy-Line Sonia Literature'. 31 May 2024. https://www.hyline.com/literature/sonia.

Iannotti, Lora L, Chessa K Lutter, David A Bunn, and Christine P Stewart. 2014. 'Eggs: The Uncracked Potential for Improving Maternal and Young Child Nutrition among the World's Poor'. *Nutrition Reviews* 72 (6): 355–68. https://doi.org/10.1111/nure.12107.

Ito, Diogo. 2023. 'Controlling Costs Through the Use of Alternative Ingredients in Poultry Diets - Laying Hens'. Summer 2023. https://layinghens.hendrix-genetics.com/en/articles/controlling-costs-through-the-use-of-alternative-ingredients-in-poultry-diets/.

Jacoby, Richard, Manuela Peukert, Antonella Succurro, Anna Koprivova, and Stanislav Kopriva. 2017. 'The Role of Soil Microorganisms in Plant Mineral Nutrition—Current Knowledge and Future Directions'. *Frontiers in Plant Science* 8 (September). https://doi.org/10.3389/fpls.2017.01617.

Jansseune, Samuel C. G., Aart Lammers, Jürgen van Baal, Fany Blanc, Marie-Hélène Pinard van der Laan, Fanny Calenge, and Wouter H. Hendriks. 2024. 'Diet Composition Influences Probiotic and Postbiotic Effects on Broiler Growth and Physiology'. *Poultry Science* 103 (6): 103650. https://doi.org/10.1016/j.psj.2024.103650.

Jehl, Frédéric. 2020. 'Étude de la composante génétique de l'efficience alimentaire (EA) chez des lignées de poules pondeuses divergentes pour l'EA en utilisant la technologie RNA-seq'. Phdthesis, Agrocampus Ouest. https://theses.hal.science/tel-03462775.

Jha, Rajesh, Janelle M. Fouhse, Utsav P. Tiwari, Linge Li, and Benjamin P. Willing. 2019. 'Dietary Fiber and Intestinal Health of Monogastric Animals'. *Frontiers in Veterinary Science* 6 (March). https://doi.org/10.3389/fvets.2019.00048.

Ji, Zhicheng, and Li Ma. 2023. 'Controlling Taxa Abundance Improves Metatranscriptomics Differential Analysis'. *BMC Microbiology* 23 (1): 60. https://doi.org/10.1186/s12866-023-02799-9.

Johnson, Jethro S., Daniel J. Spakowicz, Bo-Young Hong, Lauren M. Petersen, Patrick Demkowicz, Lei Chen, Shana R. Leopold, et al. 2019. 'Evaluation of 16S rRNA Gene Sequencing for Species and Strain-Level Microbiome Analysis'. *Nature Communications* 10 (1): 5029. https://doi.org/10.1038/s41467-019-13036-1.

Józefiak, D, A Rutkowski, and S. A Martin. 2004. 'Carbohydrate Fermentation in the Avian Ceca: A Review'. *Animal Feed Science and Technology* 113 (1): 1–15. https://doi.org/10.1016/j.anifeedsci.2003.09.007.

Kalvari, Ioanna, Eric P Nawrocki, Nancy Ontiveros-Palacios, Joanna Argasinska, Kevin Lamkiewicz, Manja Marz, Sam Griffiths-Jones, et al. 2021. 'Rfam 14: Expanded Coverage of Metagenomic, Viral and microRNA Families'. *Nucleic Acids Research* 49 (D1): D192–200. https://doi.org/10.1093/nar/gkaa1047.

Kanehisa Laboratories. n.d. 'KEGG PATHWAY: ABC Transporters'. Accessed 19 August 2024. https://www.kegg.jp/pathway/ko02010.

Kanehisa, Minoru, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. 2017. 'KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs'. *Nucleic Acids Research* 45 (D1): D353–61. https://doi.org/10.1093/nar/gkw1092.

Kanehisa, Minoru, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2016. 'KEGG as a Reference Resource for Gene and Protein Annotation'. *Nucleic Acids Research* 44 (D1): D457–62. https://doi.org/10.1093/nar/gkv1070.

Kang, Dongwan D., Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. 2019. 'MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies'. *PeerJ* 7 (July). https://doi.org/10.7717/peerj.7359.

Kers, Jannigje G., Francisca C. Velkers, Egil A. J. Fischer, Gerben D. A. Hermes, J. A. Stegeman, and Hauke Smidt. 2018. 'Host and Environmental Factors Affecting the Intestinal Microbiota in Chickens'. *Frontiers in Microbiology* 9. https://doi.org/10.3389/fmicb.2018.00235.

Khan, Samiullah, Robert J. Moore, Dragana Stanley, and Kapil K. Chousalkar. 2020. 'The Gut Microbiota of Laying Hens and Its Manipulation with Prebiotics and Probiotics To Enhance Gut Health and Food Safety'. *Applied and Environmental Microbiology* 86 (13). https://doi.org/10.1128/AEM.00600-20.

Khasanah, Himmatul, Dwi E. Kusbianto, Listya Purnamasari, Joseph F. dela Cruz, Desy C. Widianingrum, and Seong Gu Hwang. 2024. 'Modulation of Chicken Gut Microbiota for Enhanced Productivity and Health: A Review'. *Veterinary World* 17 (5): 1073–83. https://doi.org/10.14202/vetworld.2024.1073-1083.

Kim, Chan Yeong, Muyoung Lee, Sunmo Yang, Kyungnam Kim, Dongeun Yong, Hye Ryun Kim, and Insuk Lee. 2021. 'Human Reference Gut Microbiome Catalog Including Newly Assembled Genomes from Under-Represented Asian Metagenomes'. *Genome Medicine* 13 (1): 134. https://doi.org/10.1186/s13073-021-00950-7.

Kircher, Berenike, Sabrina Woltemate, Frank Gutzki, Dirk Schlüter, Robert Geffers, Heike Bähre, and Marius Vital. 2022. 'Predicting Butyrate- and Propionate-Forming Bacteria of Gut Microbiota from Sequencing Data'. *Gut Microbes* 14 (1): 2149019. https://doi.org/10.1080/19490976.2022.2149019.

Klingenberg, Heiner, and Peter Meinicke. 2017. 'How to Normalize Metatranscriptomic Count Data for Differential Expression Analysis'. *PeerJ* 5 (October):e3859. https://doi.org/10.7717/peerj.3859.

Knight, R., Alison Vrbanac, Bryn C. Taylor, A. Aksenov, C. Callewaert, Justine W. Debelius, Antonio Gonzalez, et al. 2018. 'Best Practices for Analysing Microbiomes'. *Nature Reviews Microbiology* 16:410–22. https://doi.org/10.1038/s41579-018-0029-9.

Koch, Robert. 1881. 'Zur Untersuchung von Pathogenen Organismen'. *Mittheilungen Aus Dem Kaiserlichen Gesundheitsamte* 1:1–48.

Koch, Robert, L. A. Swiger, Doyle Chambers, and K. E. Gregory. 1963. 'Efficiency of Feed Use in Beef Cattle'. *Journal of Animal Science* 22 (2): 486–94. https://doi.org/10.2527/jas1963.222486x.

Kogut, Michael H. 2022. 'Role of Diet-Microbiota Interactions in Precision Nutrition of the Chicken: Facts, Gaps, and New Concepts'. *Poultry Science* 101 (3): 101673. https://doi.org/10.1016/j.psj.2021.101673.

Kogut, Michael H., Annah Lee, and Elizabeth Santin. 2020. 'Microbiome and Pathogen Interaction with the Immune System'. *Poultry Science* 99 (4): 1906–13. https://doi.org/10.1016/j.psj.2019.12.011.

Konstantinidis, Konstantinos T., and James M. Tiedje. 2005. 'Genomic Insights That Advance the Species Definition for Prokaryotes'. *Proceedings of the National Academy of Sciences* 102 (7): 2567–72. https://doi.org/10.1073/pnas.0409727102.

Kopylova, Evguenia, Laurent Noé, and Hélène Touzet. 2012. 'SortMeRNA: Fast and Accurate Filtering of Ribosomal RNAs in Metatranscriptomic Data'. *Bioinformatics* 28 (24): 3211–17. https://doi.org/10.1093/bioinformatics/bts611.

Korver, D.R. 2023. 'Review: Current Challenges in Poultry Nutrition, Health, and Welfare'. *Animal* 17 (June):100755. https://doi.org/10.1016/j.animal.2023.100755.

Kraimi, Narjis, Ludovic Calandreau, Olivier Zemb, Karine Germain, Christèle Dupont, Philippe Velge, Edouard Guitton, Sébastien Lavillatte, Céline Parias, and Christine Leterrier. 2019. 'Effects of Gut Microbiota Transfer on Emotional Reactivity in Japanese Quails (Coturnix Japonica)'. *Journal of Experimental Biology* 222 (10): jeb202879. https://doi.org/10.1242/jeb.202879.

Kraimi, Narjis, Marian Dawkins, Sabine G. Gebhardt-Henrich, Philippe Velge, Ivan Rychlik, Jiří Volf, Pauline Creach, Adrian Smith, Frances Colles, and Christine Leterrier. 2019. 'Influence of the Microbiota-Gut-Brain Axis on Behavior and Welfare in Farm Animals: A Review'. *Physiology & Behavior* 210 (October):112658. https://doi.org/10.1016/j.physbeh.2019.112658.

Kromhout, D., C. J. K. Spaaij, J. de Goede, and R. M. Weggemans. 2016. 'The 2015 Dutch Food-Based Dietary Guidelines'. *European Journal of Clinical Nutrition* 70 (8): 869–78. https://doi.org/10.1038/ejcn.2016.52.

Labroue, F., L. Maignel, Pierre Sellier, and Jean Noblet. 1999. 'Consommation résiduelle chez le porc en croissance alimenté à volonté. Méthode de calcul et variabilité génétique'. *Journees de la Recherche Porcine en France* 31:167.

Larzul, Catherine, Jordi Estellé, Marion Borey, Fany Blanc, Gaëtan Lemonnier, Yvon Billon, Mamadou Gabou Thiam, et al. 2024. 'Driving Gut Microbiota Enterotypes through Host Genetics'. *Microbiome* 12 (1): 116. https://doi.org/10.1186/s40168-024-01827-8.

Lawal, R. A., and O. Hanotte. 2021. 'Domestic Chicken Diversity: Origin, Distribution, and Adaptation'. *Animal Genetics* 52 (4): 385–94. https://doi.org/10.1111/age.13091.

Leclercq, B, Y Henry, and F Lebas. 1996. 'Evolution de la nutrition des espèces monogastriques'.

Lederberg, Joshua, and Alexa T McCray. 2001. ''Ome Sweet 'Omics-- A Genealogical Treasury of Words'. *The Scientist* 15 (7): 8.

Legendre, Pierre, and Louis Legendre. 2012. 'Chapter 11 - Canonical Analysis'. In *Developments in Environmental Modelling*, edited by Pierre Legendre and Louis Legendre, 24:625–710. Numerical Ecology. Elsevier. https://doi.org/10.1016/B978-0-444-53868-0.50011-3.

Lehr, Alexa. 2021. 'Hens Need Calcium - Here's How to Make Sure They Get Enough | Grubbly'. Grubbly Farms. 14 May 2021. https://grubblyfarms.com/blogs/the-flyer/calcium-needs-for-hens.

Lei, Fang, Yeshi Yin, Yuezhu Wang, Bo Deng, Hongwei David Yu, Lanjuan Li, Charlie Xiang, Shengyue Wang, Baoli Zhu, and Xin Wang. 2012. 'Higher-Level Production of Volatile Fatty Acids In Vitro by Chicken Gut Microbiotas than by Human Gut Microbiotas as Determined by Functional Analyses'. *Applied and Environmental Microbiology* 78 (16): 5763–72. https://doi.org/10.1128/AEM.00327-12.

Leinonen, Ilkka, and Ilias Kyriazakis. 2016. 'How Can We Improve the Environmental Sustainability of Poultry Production?' *The Proceedings of the Nutrition Society* 75 (3): 265–73. https://doi.org/10.1017/S0029665116000094.

Lennon, J. T., M. E. Muscarella, S. A. Placella, and B. K. Lehmkuhl. 2018. 'How, When, and Where Relic DNA Affects Microbial Diversity'. *mBio* 9 (3): 10.1128/mbio.00637-18. https://doi.org/10.1128/mbio.00637-18.

Lenth, Russell V., Paul Buerkner, Maxime Herve, Jonathon Love, Fernando Miguez, Hannes Riebl, and Henrik Singmann. 2022. 'Emmeans: Estimated Marginal Means, Aka Least-Squares Means'. https://CRAN.R-project.org/package=emmeans.

Leone, Francesca, and Valentina Ferrante. 2023. 'Effects of Prebiotics and Precision Biotics on Performance, Animal Welfare and Environmental Impact. A Review'. *Science of The Total Environment* 901 (November):165951. https://doi.org/10.1016/j.scitotenv.2023.165951.

Leray, Matthieu, Nancy Knowlton, and Ryuji J. Machida. 2022. 'MIDORI2: A Collection of Quality Controlled, Preformatted, and Regularly Updated Reference Databases for Taxonomic Assignment of Eukaryotic Mitochondrial Sequences'. *Environmental DNA* 4 (4): 894–907. https://doi.org/10.1002/edn3.303.

Letunic, Ivica, and Peer Bork. 2018. '20 Years of the SMART Protein Domain Annotation Resource'. *Nucleic Acids Research* 46 (D1): D493–96. https://doi.org/10.1093/nar/gkx922.

Levasseur, Anthony, Elodie Drula, Vincent Lombard, Pedro M. Coutinho, and Bernard Henrissat. 2013. 'Expansion of the Enzymatic Repertoire of the CAZy Database to Integrate Auxiliary Redox Enzymes'. *Biotechnology for Biofuels* 6 (1): 41. https://doi.org/10.1186/1754-6834-6-41.

Li, Dinghua, Ruibang Luo, Chi-Man Liu, Chi-Ming Leung, Hing-Fung Ting, Kunihiko Sadakane, Hiroshi Yamashita, and Tak-Wah Lam. 2016. 'MEGAHIT v1.0: A Fast and Scalable Metagenome Assembler Driven by Advanced Methodologies and Community Practices'. *Methods* 102 (June):3–11. https://doi.org/10.1016/j.ymeth.2016.02.020.

Li, Fuyong, Thomas C. A. Hitch, Yanhong Chen, Christopher J. Creevey, and Le Luo Guan. 2019. 'Comparative Metagenomic and Metatranscriptomic Analyses Reveal the Breed Effect on the Rumen Microbiome and Its Associations with Feed Efficiency in Beef Cattle'. *Microbiome* 7 (1): 6. https://doi.org/10.1186/s40168-019-0618-5.

Liu, Jianan, Ying Bai, Fang Liu, Richard A. Kohn, Daniel A. Tadesse, Saul Sarria, Robert W. Li, and Jiuzhou Song. 2022. 'Rumen Microbial Predictors for Short-Chain Fatty Acid Levels and the Grass-Fed Regimen in Angus Cattle'. *Animals : An Open Access Journal from MDPI* 12 (21): 2995. https://doi.org/10.3390/ani12212995.

lohmann. 2016. 'Outstanding Performances: LOHMANN Layers above the Standards'. *Lohmann Breeders* (blog). February 2016. https://lohmann-breeders.com/outstanding-performances-lohmann-layers-above-the-standards/.

Lombard, Vincent, Hemalatha Golaconda Ramulu, Elodie Drula, Pedro M. Coutinho, and Bernard Henrissat. 2014. 'The Carbohydrate-Active Enzymes Database (CAZy) in 2013'. *Nucleic Acids Research* 42 (Database issue): D490-495. https://doi.org/10.1093/nar/gkt1178.

Louis, Petra, and Harry J. Flint. 2017. 'Formation of Propionate and Butyrate by the Human Colonic Microbiota'. *Environmental Microbiology* 19 (1): 29–41. https://doi.org/10.1111/1462-2920.13589.

Ludwig, Wolfgang, Tomeu Viver, Ralf Westram, Juan Francisco Gago, Esteban Bustos-Caparros, Katrin Knittel, Rudolf Amann, and Ramon Rossello-Mora. 2021. 'Release LTP_12_2020, Featuring a New ARB Alignment and Improved 16S rRNA Tree for Prokaryotic Type Strains'. *Systematic and Applied Microbiology* 44 (4): 126218. https://doi.org/10.1016/j.syapm.2021.126218.

Luo, Yuheng, Hong Chen, Bing Yu, Jun He, Ping Zheng, Xiangbing Mao, Gang Tian, et al. 2017. 'Dietary Pea Fiber Increases Diversity of Colonic Methanogens of Pigs with a Shift from Methanobrevibacter to Methanomassiliicoccus-like Genus and Change in Numbers of Three Hydrogenotrophs'. *BMC Microbiology* 17 (1): 17. https://doi.org/10.1186/s12866-016-0919-9.

Macelline, Shemil P., Mehdi Toghyani, Peter V. Chrystal, Peter H. Selle, and Sonia Yun Liu. 2021. 'Amino Acid Requirements for Laying Hens: A Comprehensive Review'. *Poultry Science* 100 (5): 101036. https://doi.org/10.1016/j.psj.2021.101036.

Mahé, Frédéric, Torbjørn Rognes, Christopher Quince, Colomban de Vargas, and Micah Dunthorn. 2014. 'Swarm: Robust and Fast Clustering Method for Amplicon-Based Studies'. *PeerJ* 2 (September):e593. https://doi.org/10.7717/peerj.593.

Mainguy, Jean, Adrien Castinel, Olivier Bouchez, Sylvie Combes, Carole Iampietro, Christine Gaspin, Denis Milan, Cécile Donnadieu, Claire Hoede, and Géraldine Pascal. 2022. 'Strengths and Limits of Long Read Metabarcoding'. In . Meliá Sitges, Spain.

Mainguy, Jean, and Claire Hoede. 2024. 'Binette: A Fast and Accurate Bin Refinement Tool to Construct High Quality Metagenome Assembled Genomes'. bioRxiv. https://doi.org/10.1101/2024.04.20.585171.

Manor, O., and E. Borenstein. 2017. 'Revised Computational Metagenomic Processing Uncovers Hidden and Biologically Meaningful Functional Variation in the Human Microbiome'. *Microbiome* 5 (1). https://doi.org/10.1186/s40168-017-0231-4.

Mariadassou, Mahendra, Marie Suez, Sanbadam Sathyakumar, Alain Vignal, Mariangela Arca, Pierre Nicolas, Thomas Faraut, et al. 2021. 'Unraveling the History of the Genus Gallus through Whole Genome Sequencing'. *Molecular Phylogenetics and Evolution* 158 (May):107044. https://doi.org/10.1016/j.ympev.2020.107044.

Marty, J. 1984. 'Absorption and Metabolism of the Volatile Fatty Acids in the Hind-Gut of the Rabbit'. *British Journal of Nutrition* 51 (2): 265–77. https://doi.org/10.1079/BJN19840031.

Massana, Ramon, and David López-Escardó. 2022. 'Metagenome Assembled Genomes Are for Eukaryotes Too'. *Cell Genomics* 2 (5): 100130. https://doi.org/10.1016/j.xgen.2022.100130.

McAllister, T. A., L. Dunière, P. Drouin, S. Xu, Y. Wang, K. Munns, and R. Zaheer. 2018. '*Silage Review:* Using Molecular Approaches to Define the Microbial Ecology of Silage'. *Journal of Dairy Science* 101 (5): 4060–74. https://doi.org/10.3168/jds.2017-13704.

McNeil, N.I. 1984. 'The Contribution of the Large Intestine to Energy Supplies in Man'. *The American Journal of Clinical Nutrition* 39 (2): 338–42. https://doi.org/10.1093/ajcn/39.2.338.

Mehta, Raaj S., Galeb S. Abu-Ali, David A. Drew, Jason Lloyd-Price, Ayshwarya Subramanian, Paul Lochhead, Amit D. Joshi, et al. 2018. 'Stability of the Human Faecal Microbiome in a Cohort of Adult Men'. *Nature Microbiology* 3 (3): 347–55. https://doi.org/10.1038/s41564-017-0096-0.

Meola, Marco, Etienne Rifa, Noam Shani, Céline Delbès, Hélène Berthoud, and Christophe Chassard. 2019. 'DAIRYdb: A Manually Curated Reference Database for Improved Taxonomy Annotation

of 16S rRNA Gene Sequences from Dairy Products'. *BMC Genomics* 20 (1): 560. https://doi.org/10.1186/s12864-019-5914-8.

Metzler-Zebeli, Barbara U., Sina-Catherine Siegerstetter, Elizabeth Magowan, Peadar G. Lawlor, Renée M. Petri, Niamh E. O´Connell, and Qendrim Zebeli. 2019. 'Feed Restriction Modifies Intestinal Microbiota-Host Mucosal Networking in Chickens Divergent in Residual Feed Intake'. *mSystems* 4 (1): e00261-18. https://doi.org/10.1128/mSystems.00261-18.

Mignon-Grasteau, Sandrine, Agnès Narcy, Nicole Rideau, Céline Chantry-Darmon, Marie-Yvonne Boscher, Nadine Sellier, Marie Chabault, Barbara Konsak-Ilievski, Elisabeth Le Bihan-Duval, and Irène Gabriel. 2015. 'Impact of Selection for Digestive Efficiency on Microbiota Composition in the Chicken'. *PLOS ONE* 10 (8): e0135488. https://doi.org/10.1371/journal.pone.0135488.

Misiukiewicz, A., M. Gao, W. Filipiak, A. Cieslak, A. K. Patra, and M. Szumacher-Strabel. 2021. 'Review: Methanogens and Methane Production in the Digestive Systems of Nonruminant Farm Animals'. *Animal* 15 (1): 100060. https://doi.org/10.1016/j.animal.2020.100060.

Mittal, Rahul, Sebastian V. Sanchez-Luege, Shannon Michele Wagner, Denise Yan, and Xue Zhong Liu. 2019. 'Recent Perspectives on Gene-Microbe Interactions Determining Predisposition to Otitis Media'. *Frontiers in Genetics* 10. https://doi.org/10.3389/fgene.2019.01230.

Moinard, Sylvain, Didier Piau, Frédéric Laporte, Delphine Rioux, Pierre Taberlet, Christelle Gonindard-Melodelima, and Eric Coissac. 2023. 'Towards Quantitative DNA Metabarcoding: A Method to Overcome PCR Amplification Bias'. bioRxiv. https://doi.org/10.1101/2023.10.03.560640.

Moissl-Eichinger, Christine, Manuela Pausan, Julian Taffner, Gabriele Berg, Corinna Bang, and Ruth A. Schmitz. 2018. 'Archaea Are Interactive Components of Complex Microbiomes'. *Trends in Microbiology* 26 (1): 70–85. https://doi.org/10.1016/j.tim.2017.07.004.

Morrison, Douglas J., and Tom Preston. 2016. 'Formation of Short Chain Fatty Acids by the Gut Microbiota and Their Impact on Human Metabolism'. *Gut Microbes* 7 (3): 189–200. https://doi.org/10.1080/19490976.2015.1134082.

Nadkarni, Mangala A, F. Elizabeth Martin, Nicholas A Jacques, and Neil Hunter. 2002. 'Determination of Bacterial Load by Real-Time PCR Using a Broad-Range (Universal) Probe and Primers Set'. *Microbiology* 148 (1): 257–66. https://doi.org/10.1099/00221287-148-1-257.

NCBI Staff. 2021. 'NCBI Taxonomy to Include Phylum Rank in Taxonomic Names'. NCBI Insights. 10 December 2021. https://ncbiinsights.ncbi.nlm.nih.gov/2021/12/10/ncbi-taxonomy-prokaryote-phyla-added/.

———. 2022. 'Prokaryotic Phylum Name Changes Coming Soon!' NCBI Insights. 14 November 2022. https://ncbiinsights.ncbi.nlm.nih.gov/2022/11/14/prokaryotic-phylum-name-changes/.

———. 2023. 'Upcoming Changes to Influenza Virus Names in NCBI Taxonomy'. NCBI Insights. 21 February 2023. https://ncbiinsights.ncbi.nlm.nih.gov/2023/02/21/influenza-virus-ncbi-taxonomy/.

———. n.d. 'Major 3rd Party Taxonomic Data Resources'. Accessed 16 July 2024. https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=resources.

Neeteson-van Nieuwenhoven, Anne-Marie, Pieter Knap, and Santiago Avendaño. 2013. 'The Role of Sustainable Commercial Pig and Poultry Breeding for Food Security'. *Animal Frontiers* 3 (1): 52–57. https://doi.org/10.2527/af.2013-0008.

Nielsen, H. Bjørn, Mathieu Almeida, Agnieszka Sierakowska Juncker, Simon Rasmussen, Junhua Li, Shinichi Sunagawa, Damian R. Plichta, et al. 2014. 'Identification and Assembly of Genomes and Genetic Elements in Complex Metagenomic Samples without Using Reference Genomes'. *Nature Biotechnology* 32 (8): 822–28. https://doi.org/10.1038/nbt.2939.

Núñez-Sánchez, María A., Silvia Melgar, Keith O'Donoghue, María A. Martínez-Sánchez, Virgina E. Fernández-Ruiz, Mercedes Ferrer-Gómez, Antonio J. Ruiz-Alcaraz, and Bruno Ramos-Molina. 2022. 'Crohn's Disease, Host–Microbiota Interactions, and Immunonutrition: Dietary Strategies Targeting Gut Microbiome as Novel Therapeutic Approaches'. *International Journal of Molecular Sciences* 23 (15): 8361. https://doi.org/10.3390/ijms23158361.

Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner. 2017. 'metaSPAdes: A New Versatile Metagenomic Assembler'. *Genome Research* 27 (5): 824–34. https://doi.org/10.1101/gr.213959.116.

Nys, Y., and N. Guyot. 2011. 'Egg Formation and Chemistry'. In *Improving the Safety and Quality of Eggs and Egg Products*, 83–132. Elsevier. https://doi.org/10.1533/9780857093912.2.83.

Oakley, Brian B., Hyun S. Lillehoj, Michael H. Kogut, Woo K. Kim, John J. Maurer, Adriana Pedroso, Margie D. Lee, Stephen R. Collett, Timothy J. Johnson, and Nelson A. Cox. 2014. 'The Chicken Gastrointestinal Microbiome'. *FEMS Microbiology Letters* 360 (2): 100–112. https://doi.org/10.1111/1574-6968.12608.

Ogier, Jean-Claude, Sylvie Pagès, Maxime Galan, Matthieu Barret, and Sophie Gaudriault. 2019. 'rpoB, a Promising Marker for Analyzing the Diversity of Bacterial Communities by Amplicon Sequencing'. *BMC Microbiology* 19 (1): 171. https://doi.org/10.1186/s12866-019-1546-z.

Oksanen, Jari, Gavin L. Simpson, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, et al. 2022. 'Vegan: Community Ecology Package'. https://CRAN.R-project.org/package=vegan.

Pan, Deng, and Zhongtang Yu. 2014. 'Intestinal Microbiome of Poultry and Its Interaction with Host and Diet'. *Gut Microbes* 5 (1): 108–19. https://doi.org/10.4161/gmic.26945.

Parks, Donovan H., Maria Chuvochina, Christian Rinke, Aaron J. Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2022. 'GTDB: An Ongoing Census of Bacterial and Archaeal Diversity through a Phylogenetically Consistent, Rank Normalized and Complete Genome-Based Taxonomy'. *Nucleic Acids Research* 50 (D1): D785–94. https://doi.org/10.1093/nar/gkab776.

Parte, Aidan C. 2018. 'LPSN – List of Prokaryotic Names with Standing in Nomenclature (Bacterio.Net), 20 Years On'. *International Journal of Systematic and Evolutionary Microbiology* 68 (6): 1825–29. https://doi.org/10.1099/ijsem.0.002786.

Pascal, Géraldine. 2023. '16S-ITS-23S-DB database'. Recherche Data Gouv. https://doi.org/10.57745/APDPLQ.

Pauwels, J., B. Taminiau, G. P. J. Janssens, M. De Beenhouwer, L. Delhalle, G. Daube, and F. Coopman. 2015. 'Cecal Drop Reflects the Chickens' Cecal Microbiome, Fecal Drop Does Not'. *Journal of Microbiological Methods* 117 (October):164–70. https://doi.org/10.1016/j.mimet.2015.08.006.

Peng, Lu-Yuan, Hai-Tao Shi, Zi-Xuan Gong, Peng-Fei Yi, Bo Tang, Hai-Qing Shen, and Ben-Dong Fu. 2021. 'Protective Effects of Gut Microbiota and Gut Microbiota-Derived Acetate on Chicken Colibacillosis Induced by Avian Pathogenic Escherichia Coli'. *Veterinary Microbiology* 261 (October):109187. https://doi.org/10.1016/j.vetmic.2021.109187.

Pesant, Stéphane, Fabrice Not, Marc Picheral, Stefanie Kandels-Lewis, Noan Le Bescot, Gabriel Gorsky, Daniele Iudicone, et al. 2015. 'Open Science Resources for the Discovery and Analysis of Tara Oceans Data'. *Scientific Data* 2 (1): 150023. https://doi.org/10.1038/sdata.2015.23.

Pimentel, Mark, and Anthony Lembo. 2020. 'Microbiome and Its Role in Irritable Bowel Syndrome'. *Digestive Diseases and Sciences* 65 (3): 829–39. https://doi.org/10.1007/s10620-020-06109-5.

Plancade, Sandra, Allison Clark, Catherine Philippe, Jean-Christophe Helbling, Marie-Pierre Moisan, Diane Esquerré, Laurence Le Moyec, Céline Robert, Eric Barrey, and Núria Mach. 2019. 'Unraveling the Effects of the Gut Microbiota Composition and Function on Horse Endurance Physiology'. *Scientific Reports* 9 (1): 9620. https://doi.org/10.1038/s41598-019-46118-7.

Plaza Oñate, Florian, Marie Jeammet, Nicolas pons, Stanislav Dusko Ehrlich, Jordi Estellé, and Fanny Calenge. 2023. 'MetaChick: characterization of the chicken caecal metagenome by deep shotgun sequencing'. Recherche Data Gouv. https://doi.org/10.15454/FHPJH5.

Plaza Oñate, Florian, Emmanuelle Le Chatelier, Mathieu Almeida, Alessandra C L Cervino, Franck Gauthier, Frédéric Magoulès, S Dusko Ehrlich, and Matthieu Pichaud. 2019. 'MSPminer: Abundance-Based Reconstitution of Microbial Pan-Genomes from Shotgun Metagenomic Data'. *Bioinformatics* 35 (9): 1544–52. https://doi.org/10.1093/bioinformatics/bty830.

Pons, Nicolas, Jean-Michel Batto, Sean Kennedy, Mathieu Almeida, Fouad Boumezbeur, Bouziane Moumen, and Pierre Leonard. 2010. 'METEOR - a Plateform for Quantitative Metagenomic Profiling of Complex Ecosystems'. In *JOBIM*. Montpellier, France.

Porter, Teresita M., and Mehrdad Hajibabaei. 2022. 'MetaWorks: A Flexible, Scalable Bioinformatic Pipeline for High-Throughput Multi-Marker Biodiversity Assessments'. *PLOS ONE* 17 (9): e0274260. https://doi.org/10.1371/journal.pone.0274260.

Pradeep. 2020. 'The Ultimate Guide to FCR (Feed Conversion Ratio) - Navfarm Blog'. *Navfarm* (blog). 22 September 2020. https://www.navfarm.com/blog/fcr-guide/.

Preisinger, Rudolf. 2018. 'Innovative Layer Genetics to Handle Global Challenges in Egg Production'. *British Poultry Science* 59 (1): 1–6. https://doi.org/10.1080/00071668.2018.1401828.

Pym, R A E, E Guerne Bleich, and I Hoffmann. 2006. 'The Relative Contribution of Indigenous Chicken Breeds to Poultry Meat and Egg Production and Consumption in the Developing Countries of Africa and Asia.' In . Italy: World's Poultry Science Assoc.

Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, et al. 2010. 'A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing'. *Nature* 464 (7285): 59–65. https://doi.org/10.1038/nature08821.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2013. 'The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools'. *Nucleic Acids Research* 41 (Database issue): D590–96. https://doi.org/10.1093/nar/gks1219.

Quéré, Pascale. 2017. 'Contrasting Effects of Genetic Selection on Feed Efficiency on Resistance to Avian Influenza and Colibacillosis in the Chicken'. Poster presented at the The Avian Genetics and Immunity Symposium, Guildford United Kingdom, June.

Quince, Christopher, Alan W. Walker, Jared T. Simpson, Nicholas J. Loman, and Nicola Segata. 2017. 'Shotgun Metagenomics, from Sampling to Analysis'. *Nature Biotechnology* 35 (9): 833–44. https://doi.org/10.1038/nbt.3935.

Rao, C Radhakrishna. 1964. 'The Use and Interpretation of Principal Component Analysis in Applied Research'. *Sankhyā: The Indian Journal of Statistics,* Series A (1961-2002), 26 (4): 329–58.

Rausch, Philipp, Malte Rühlemann, Britt M. Hermes, Shauni Doms, Tal Dagan, Katja Dierking, Hanna Domin, et al. 2019. 'Comparative Analysis of Amplicon and Metagenomic Sequencing Methods Reveals Key Features in the Evolution of Animal Metaorganisms'. *Microbiome* 7 (1): 133. https://doi.org/10.1186/s40168-019-0743-1.

Ravindran, Velmurugu. 2012. 'Advances and Future Directions in Poultry Nutrition: An Overview'. *Korean Journal of Poultry Science* 39 (1): 53–62. https://doi.org/10.5536/KJPS.2012.39.1.053.

———. 2013. 'Poultry Feed Availability and Nutrition in Developing Countries'. *FAO*. https://www.fao.org/3/i3531e/i3531e.pdf.

Réhault-Godbert, Sophie, Nicolas Guyot, and Yves Nys. 2019. 'The Golden Egg: Nutritional Value, Bioactivities, and Emerging Benefits for Human Health'. *Nutrients* 11 (3): 684. https://doi.org/10.3390/nu11030684.

Reichardt, Nicole, Sylvia H Duncan, Pauline Young, Alvaro Belenguer, Carol McWilliam Leitch, Karen P Scott, Harry J Flint, and Petra Louis. 2014. 'Phylogenetic Distribution of Three Pathways for Propionate Production within the Human Gut Microbiota'. *The ISME Journal* 8 (6): 1323–35. https://doi.org/10.1038/ismej.2014.14.

Richards, Peter, Jo Fothergill, Marion Bernardeau, and Paul Wigley. 2019. 'Development of the Caecal Microbiota in Three Broiler Breeds'. *Frontiers in Veterinary Science* 6 (June). https://doi.org/10.3389/fvets.2019.00201.

Rimet, Frédéric, Evgenuy Gusev, Maria Kahlert, Martyn G. Kelly, Maxim Kulikovskiy, Yevhen Maltsev, David G. Mann, et al. 2019. 'Diat.Barcode, an Open-Access Curated Barcode Library for Diatoms'. *Scientific Reports* 9 (1): 15116. https://doi.org/10.1038/s41598-019-51500-6.

Rintala, Anniina, Sami Pietilä, Eveliina Munukka, Erkki Eerola, Juha-Pekka Pursiheimo, Asta Laiho, Satu Pekkala, and Pentti Huovinen. 2017. 'Gut Microbiota Analysis Results Are Highly Dependent

on the 16S rRNA Gene Target Region, Whereas the Impact of DNA Extraction Is Minor'. *Journal of Biomolecular Techniques : JBT* 28 (1): 19–30. https://doi.org/10.7171/jbt.17-2801-003.

Ritchie, Hannah, Pablo Rosado, and Max Roser. 2024. 'Meat and Dairy Production'. https://ourworldindata.org/meat-production.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015a. 'Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies'. *Nucleic Acids Research* 43 (7): e47. https://doi.org/10.1093/nar/gkv007.

———. 2015b. 'Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies'. *Nucleic Acids Research* 43 (7): e47. https://doi.org/10.1093/nar/gkv007.

Rivière, Audrey, Marija Selak, David Lantin, Frédéric Leroy, and Luc De Vuyst. 2016. 'Bifidobacteria and Butyrate-Producing Colon Bacteria: Importance and Strategies for Their Stimulation in the Human Gut'. *Frontiers in Microbiology* 7. https://doi.org/10.3389/fmicb.2016.00979.

Robert, Vincent, Duong Vu, Ammar Ben Hadj Amor, Nathalie van de Wiele, Carlo Brouwer, Bernard Jabas, Szaniszlo Szoke, et al. 2013. 'MycoBank Gearing up for New Horizons'. *IMA Fungus* 4 (2): 371–79. https://doi.org/10.5598/imafungus.2013.04.02.16.

Robinson, Kelsy, Zhuo Deng, Yongqing Hou, and Guolong Zhang. 2015. 'Regulation of the Intestinal Barrier Function by Host Defense Peptides'. *Frontiers in Veterinary Science* 2 (November). https://doi.org/10.3389/fvets.2015.00057.

Rosselló-Móra, Ramon, and Rudolf Amann. 2015. 'Past and Future Species Definitions for Bacteria and Archaea'. *Systematic and Applied Microbiology* 38 (4): 209–16. https://doi.org/10.1016/j.syapm.2015.02.001.

Rothschild, Daphna, Omer Weissbrod, Elad Barkan, Alexander Kurilshikov, Tal Korem, David Zeevi, Paul I. Costea, et al. 2018. 'Environment Dominates over Host Genetics in Shaping Human Gut Microbiota'. *Nature* 555 (7695): 210–15. https://doi.org/10.1038/nature25973.

Rychlik, Ivan. 2020. 'Composition and Function of Chicken Gut Microbiota'. *Animals : An Open Access Journal from MDPI* 10 (1): 103. https://doi.org/10.3390/ani10010103.

Sadagopan, Aishwarya, Anas Mahmoud, Maha Begg, Mawada Tarhuni, Monique Fotso, Natalie A. Gonzalez, Raghavendra R. Sanivarapu, Usama Osman, Abishek Latha Kumar, and Lubna Mohammed. 2023. 'Understanding the Role of the Gut Microbiome in Diabetes and Therapeutics Targeting Leaky Gut: A Systematic Review'. *Cureus* 15 (7): e41559. https://doi.org/10.7759/cureus.41559.

Saengkerdsub, Suwat, Robin C. Anderson, Heather H. Wilkinson, Woo-Kyun Kim, David J. Nisbet, and Steven C. Ricke. 2007. 'Identification and Quantification of Methanogenic Archaea in Adult Chicken Ceca'. *Applied and Environmental Microbiology* 73 (1): 353–56. https://doi.org/10.1128/AEM.01931-06.

Sambeek, Ir Frans van. 2010. 'Longer Production Cycles from a Genetic Perspective'. *International Poultry Production* 19 (1). http://www.positiveaction.info/pdfs/articles/pp19.1p27.pdf.

Santos, Andres, Ronny van Aerle, Leticia Barrientos, and Jaime Martinez-Urtaza. 2020. 'Computational Methods for 16S Metabarcoding Studies Using Nanopore Sequencing Data'. *Computational and Structural Biotechnology Journal* 18 (January):296–305. https://doi.org/10.1016/j.csbj.2020.01.005.

Schleifer, Karl Heinz. 2009. 'Classification of Bacteria and Archaea: Past, Present and Future'. *Systematic and Applied Microbiology* 32 (8): 533–42. https://doi.org/10.1016/j.syapm.2009.09.002.

Schloss, Patrick D., Sarah L. Westcott, Thomas Ryabin, Justine R. Hall, Martin Hartmann, Emily B. Hollister, Ryan A. Lesniewski, et al. 2009. 'Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities'. *Applied and Environmental Microbiology* 75 (23): 7537–41. https://doi.org/10.1128/AEM.01541-09.

Schoch, Conrad L, Stacy Ciufo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, et al. 2020. 'NCBI Taxonomy: A Comprehensive Update on

Curation, Resources and Tools'. *Database* 2020 (January):baaa062. https://doi.org/10.1093/database/baaa062.

Segura-Wang, Maia, Nikolaus Grabner, Andreas Koestelbauer, Viviana Klose, and Mahdi Ghanbari. 2021. 'Genome-Resolved Metagenomics of the Chicken Gut Microbiome'. *Frontiers in Microbiology* 12. https://www.frontiersin.org/article/10.3389/fmicb.2021.726923.

Sekelja, M., I. Rud, S. H. Knutsen, V. Denstadli, B. Westereng, T. Næs, and K. Rudi. 2012. 'Abrupt Temporal Fluctuations in the Chicken Fecal Microbiota Are Explained by Its Gastrointestinal Origin'. *Applied and Environmental Microbiology* 78 (8): 2941–48. https://doi.org/10.1128/AEM.05391-11.

Selle, PeterH., and Sonia Yun Liu. 2019. 'The Relevance of Starch and Protein Digestive Dynamics in Poultry'. *Journal of Applied Poultry Research* 28 (3): 531–45. https://doi.org/10.3382/japr/pfy026.

Seol, Donghyeok, Jin Soo Lim, Samsun Sung, Young Ho Lee, Misun Jeong, Seoae Cho, Woori Kwak, and Heebal Kim. 2022. 'Microbial Identification Using rRNA Operon Region: Database and Tool for Metataxonomics with Long-Read Sequence'. *Microbiology Spectrum* 10 (2): e02017-21. https://doi.org/10.1128/spectrum.02017-21.

Shang, Yue, Sanjay Kumar, Brian Oakley, and Woo Kyun Kim. 2018. 'Chicken Gut Microbiota: Importance and Detection Technology'. *Frontiers in Veterinary Science* 5. https://doi.org/10.3389/fvets.2018.00254.

Shcherbatov, Vyacheslav I., and Artem G. Shkuro. 2021. 'Cycles and Intervals in Hen Egg Laying'. Edited by S. Belousov and S. Roshchupkin. *E3S Web of Conferences* 285:04009. https://doi.org/10.1051/e3sconf/202128504009.

Sidibe, Ouléye, Anne-Carmen Sanchez, Guillaume Kon Kam King, Fanny Calenge, Benoît Doublet, Sylvie Baucheron, Sébastien Leclercq, and Anne-Laure Abraham. 2023. 'Characterization and Quantification of Antibiotic Resistance Gene Variants in Gut Microbiota.' In *JOBIM 2023*. https://hal.inrae.fr/hal-04170846.

Siegerstetter, Sina-Catherine, Renee M. Petri, Elizabeth Magowan, Peadar G. Lawlor, Qendrim Zebeli, Niamh E. O'Connell, and Barbara U. Metzler-Zebeli. 2018. 'Feed Restriction Modulates the Fecal Microbiota Composition, Nutrient Retention, and Feed Efficiency in Chickens Divergent in Residual Feed Intake'. *Frontiers in Microbiology* 9 (November):2698. https://doi.org/10.3389/fmicb.2018.02698.

Simon, Jean-Christophe, Julian R. Marchesi, Christophe Mougel, and Marc-André Selosse. 2019. 'Host-Microbiota Interactions: From Holobiont Theory to Analysis'. *Microbiome* 7 (1): 5. https://doi.org/10.1186/s40168-019-0619-4.

Sinclair-Black, Micaela, R. Alejandra Garcia, and Laura E. Ellestad. 2023. 'Physiological Regulation of Calcium and Phosphorus Utilization in Laying Hens'. *Frontiers in Physiology* 14 (February):1112499. https://doi.org/10.3389/fphys.2023.1112499.

Slavin, Joanne L., and Beate Lloyd. 2012. 'Health Benefits of Fruits and Vegetables'. *Advances in Nutrition* 3 (4): 506–16. https://doi.org/10.3945/an.112.002154.

Sokal, Robert R. 1963. 'The Principles and Practice of Numerical Taxonomy'. *TAXON* 12 (5): 190–99. https://doi.org/10.2307/1217562.

Stackebrandt, E., and J. Ebers. 2006. 'Taxonomic Parameters Revisited: Tarnished Gold Standards'. *MICROBIOLOGY TODAY* 33 (4): 152–55.

Stackebrandt, E., and B. M. Goebel. 1994. 'Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology'. *International Journal of Systematic and Evolutionary Microbiology* 44 (4): 846–49. https://doi.org/10.1099/00207713-44-4-846.

Stanley, Dragana, Mark S. Geier, Honglei Chen, Robert J. Hughes, and Robert J. Moore. 2015. 'Comparison of Fecal and Cecal Microbiotas Reveals Qualitative Similarities but Quantitative Differences'. *BMC Microbiology* 15 (February):51. https://doi.org/10.1186/s12866-015-0388-6.

Stanley, Dragana, Robert J. Hughes, and Robert J. Moore. 2014. 'Microbiota of the Chicken Gastrointestinal Tract: Influence on Health, Productivity and Disease'. *Applied Microbiology and Biotechnology* 98 (10): 4301–10. https://doi.org/10.1007/s00253-014-5646-2.

Stewart, Robert D., Marc D. Auffret, Amanda Warr, Andrew H. Wiser, Maximilian O. Press, Kyle W. Langford, Ivan Liachko, et al. 2018. 'Assembly of 913 Microbial Genomes from Metagenomic Sequencing of the Cow Rumen'. *Nature Communications* 9 (1): 870. https://doi.org/10.1038/s41467-018-03317-6.

Stoddard, Steven F., Byron J. Smith, Robert Hein, Benjamin R.K. Roller, and Thomas M. Schmidt. 2015. 'rrnDB: Improved Tools for Interpreting rRNA Gene Abundance in Bacteria and Archaea and a New Foundation for Future Development'. *Nucleic Acids Research* 43 (D1): D593–98. https://doi.org/10.1093/nar/gku1201.

Suau, A., R. Bonnet, M. Sutren, J. J. Godon, G. R. Gibson, M. D. Collins, and J. Doré. 1999. 'Direct Analysis of Genes Encoding 16S rRNA from Complex Communities Reveals Many Novel Molecular Species within the Human Gut'. *Applied and Environmental Microbiology* 65 (11): 4799–4807. https://doi.org/10.1128/AEM.65.11.4799-4807.1999.

Svihus, Birger. 2011. 'Limitations to Wheat Starch Digestion in Growing Broiler Chickens: A Brief Review'. *Animal Production Science* 51 (7): 583–89. https://doi.org/10.1071/AN10271.

Tabrett, Alexandra, and Matthew W. Horton. 2020. 'The Influence of Host Genetics on the Microbiome'. *F1000Research* 9 (February). https://doi.org/10.12688/f1000research.20835.1.

Tapio, Ilma, Timothy J. Snelling, Francesco Strozzi, and R. John Wallace. 2017. 'The Ruminal Microbiome Associated with Methane Emissions from Ruminant Livestock'. *Journal of Animal Science and Biotechnology* 8 (1): 7. https://doi.org/10.1186/s40104-017-0141-0.

Tejeda, Oscar J., and Woo K. Kim. 2021. 'Role of Dietary Fiber in Poultry Nutrition'. *Animals : An Open Access Journal from MDPI* 11 (2): 461. https://doi.org/10.3390/ani11020461.

Tessler, M., Johannes S. Neumann, Ebrahim Afshinnekoo, Michael Pineda, Rebecca Hersch, L. Velho, B. T. Segovia, et al. 2017. 'Large-Scale Differences in Microbial Biodiversity Discovery between 16S Amplicon and Shotgun Sequencing'. *Scientific Reports* 7:null. https://doi.org/10.1038/s41598-017-06665-3.

The Gene Ontology Consortium. 2017. 'Expansion of the Gene Ontology Knowledgebase and Resources'. *Nucleic Acids Research* 45 (Database issue): D331–38. https://doi.org/10.1093/nar/gkw1108.

Thioulouse, J., and D. Chessel. 1987. 'Les Analyses Multitableaux En Ecologie Factorielle.' *Acta Oecologica, Oecologia Generalis* 8 (4): 463–80.

Thomas, Christoph, and Robert Tampé. 2020. 'Structural and Mechanistic Principles of ABC Transporters'. *Annual Review of Biochemistry* 89 (Volume 89, 2020): 605–36. https://doi.org/10.1146/annurev-biochem-011520-105201.

Tito, Raúl Y., Simone Macmil, Graham Wiley, Fares Najar, Lauren Cleeland, Chunmei Qu, Ping Wang, et al. 2008. 'Phylotyping and Functional Analysis of Two Ancient Human Microbiomes'. *PloS One* 3 (11): e3703. https://doi.org/10.1371/journal.pone.0003703.

Tiwari, Utsav P., Amit K. Singh, and Rajesh Jha. 2019. 'Fermentation Characteristics of Resistant Starch, Arabinoxylan, and β-Glucan and Their Effects on the Gut Microbial Ecology of Pigs: A Review'. *Animal Nutrition* 5 (3): 217–26. https://doi.org/10.1016/j.aninu.2019.04.003.

Tixier-Boichard, Michèle, Bertrand Bed'hom, and Xavier Rognon. 2011. 'Chicken Domestication: From Archeology to Genomics'. *Comptes Rendus Biologies*, On the trail of domestications, migrations and invasions in agriculture, 334 (3): 197–204. https://doi.org/10.1016/j.crvi.2010.12.012.

Tixier-Boichard, Michèle, D. Boichard, E. Groeneveld, and A. Bordas. 1995. 'Restricted Maximum Likelihood Estimates of Genetic Parameters of Adult Male and Female Rhode Island Red Chickens Divergently Selected for Residual Feed Consumption'. *Poultry Science* 74 (8): 1245–52. https://doi.org/10.3382/ps.0741245.

Tixier-Boichard, Michèle, A Bordas, Gilles Renand, and Jean Pierre Bidanel. 2002. 'Residual Food Consumption as a Tool to Unravel Genetic Components of Food Intake'. In , 8. Montpellier, France.

Torok, Valeria A., Gwen E. Allison, Nigel J. Percy, Kathy Ophel-Keller, and Robert J. Hughes. 2011. 'Influence of Antimicrobial Feed Additives on Broiler Commensal Posthatch Gut Microbiota Development and Performance'. *Applied and Environmental Microbiology* 77 (10): 3380–90. https://doi.org/10.1128/AEM.02300-10.

Val-Laillet, David, Marie Besson, Sylvie Guérin, Nicolas Coquery, Gwénaëlle Randuineau, Ameni Kanzari, Hélène Quesnel, et al. 2017. 'A Maternal Western Diet during Gestation and Lactation Modifies Offspring's Microbiota Activity, Blood Lipid Levels, Cognitive Responses, and Hippocampal Neurogenesis in Yucatan Pigs'. *The FASEB Journal* 31 (5): 2037–49. https://doi.org/10.1096/fj.201601015R.

Vasimuddin, Md., Sanchit Misra, Heng Li, and Srinivas Aluru. 2019. 'Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems'. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, 314–24. https://doi.org/10.1109/IPDPS.2019.00041.

Velasco-Galilea, María, Miriam Piles, Yuliaxis Ramayo-Caldas, and Juan P. Sánchez. 2021. 'The Value of Gut Microbiota to Predict Feed Efficiency and Growth of Rabbits under Different Feeding Regimes'. *Scientific Reports* 11 (1): 19495. https://doi.org/10.1038/s41598-021-99028-y.

Větrovský, Tomáš, and Petr Baldrian. 2013. 'The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses'. *PLOS ONE* 8 (2): e57923. https://doi.org/10.1371/journal.pone.0057923.

Videnska, Petra, Marcela Faldynova, Helena Juricova, Vladimir Babak, Frantisek Sisak, Hana Havlickova, and Ivan Rychlik. 2013. 'Chicken Faecal Microbiota and Disturbances Induced by Single or Repeated Therapy with Tetracycline and Streptomycin'. *BMC Veterinary Research* 9 (1): 30. https://doi.org/10.1186/1746-6148-9-30.

Videnska, Petra, Karel Sedlar, Maja Lukac, Marcela Faldynova, Lenka Gerzova, Darina Cejkova, Frantisek Sisak, and Ivan Rychlik. 2014. 'Succession and Replacement of Bacterial Populations in the Caecum of Egg Laying Hens over Their Whole Life'. *PLOS ONE* 9 (12): e115142. https://doi.org/10.1371/journal.pone.0115142.

Vincent, Antony T., Nicolas Derome, Brian Boyle, Alexander I. Culley, and Steve J. Charette. 2017. 'Next-Generation Sequencing (NGS) in the Microbiological World: How to Make the Most of Your Money'. *Journal of Microbiological Methods*, What's next in microbiology methods? Emerging methods, 138 (July):60–71. https://doi.org/10.1016/j.mimet.2016.02.016.

Vogel, Timothy M., Pascal Simonet, Janet K. Jansson, Penny R. Hirsch, James M. Tiedje, Jan Dirk van Elsas, Mark J. Bailey, Renaud Nalin, and Laurent Philippot. 2009. 'TerraGenome: A Consortium for the Sequencing of a Soil Metagenome'. *Nature Reviews Microbiology* 7 (4): 252–252. https://doi.org/10.1038/nrmicro2119.

Vollmers, John, Sandra Wiegand, and Anne-Kristin Kaster. 2017. 'Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters!' *PLOS ONE* 12 (1): e0169662. https://doi.org/10.1371/journal.pone.0169662.

Wadi, Lina, Mona Meyer, Joel Weiser, Lincoln D. Stein, and Jüri Reimand. 2016. 'Impact of Outdated Gene Annotations on Pathway Enrichment Analysis'. *Nature Methods* 13 (9): 705–6. https://doi.org/10.1038/nmeth.3963.

Wagner, Josef, Paul Coupland, Hilary P. Browne, Trevor D. Lawley, Suzanna C. Francis, and Julian Parkhill. 2016. 'Evaluation of PacBio Sequencing for Full-Length Bacterial 16S rRNA Gene Classification'. *BMC Microbiology* 16 (1): 274. https://doi.org/10.1186/s12866-016-0891-4.

Wahl, Anika, Christopher Huptas, and Klaus Neuhaus. 2022. 'Comparison of rRNA Depletion Methods for Efficient Bacterial mRNA Sequencing'. *Scientific Reports* 12 (1): 5765. https://doi.org/10.1038/s41598-022-09710-y.

Walsh, Calum J., Meghana Srinivas, Timothy P. Stinear, Douwe van Sinderen, Paul D. Cotter, and John G. Kenny. 2024. 'GROND: A Quality-Checked and Publicly Available Database of Full-Length

16S-ITS-23S rRNA Operon Sequences'. *Microbial Genomics* 10 (6): 001255. https://doi.org/10.1099/mgen.0.001255.

Wang, Meng, Samina Noor, Ran Huan, Congling Liu, JiaYi Li, Qingxin Shi, Yan-Jiao Zhang, Cuiling Wu, and Hailun He. 2020. 'Comparison of the Diversity of Cultured and Total Bacterial Communities in Marine Sediment Using Culture-Dependent and Sequencing Methods'. *PeerJ* 8 (October):e10060. https://doi.org/10.7717/peerj.10060.

Wang, Qiong, George M. Garrity, James M. Tiedje, and James R. Cole. 2007. 'Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy'. *Applied and Environmental Microbiology* 73 (16): 5261–67. https://doi.org/10.1128/AEM.00062-07.

Wang, Yanan, Yongfei Hu, Fei Liu, Jian Cao, Na Lv, Baoli Zhu, Gaiping Zhang, and George Fu Gao. 2020. 'Integrated Metagenomic and Metatranscriptomic Profiling Reveals Differentially Expressed Resistomes in Human, Chicken, and Pig Gut Microbiomes'. *Environment International* 138 (May):105649. https://doi.org/10.1016/j.envint.2020.105649.

Wang, Zhi, Alexandra S. Tauzin, Elisabeth Laville, and Gabrielle Potocki-Veronese. 2022. 'Identification of Glycoside Transporters From the Human Gut Microbiome'. *Frontiers in Microbiology* 13 (March):816462. https://doi.org/10.3389/fmicb.2022.816462.

Wemheuer, Franziska, Jessica A. Taylor, Rolf Daniel, Emma Johnston, Peter Meinicke, Torsten Thomas, and Bernd Wemheuer. 2020. 'Tax4Fun2: Prediction of Habitat-Specific Functional Profiles and Functional Redundancy Based on 16S rRNA Gene Sequences'. *Environmental Microbiome* 15 (1): 11. https://doi.org/10.1186/s40793-020-00358-7.

Wen, Chaoliang, Wei Yan, Chunning Mai, Zhongyi Duan, Jiangxia Zheng, Congjiao Sun, and Ning Yang. 2021. 'Joint Contributions of the Gut Microbiota and Host Genetics to Feed Efficiency in Chickens'. *Microbiome* 9 (1): 126. https://doi.org/10.1186/s40168-021-01040-x.

Westermann, Alexander J., and Jörg Vogel. 2021. 'Cross-Species RNA-Seq for Deciphering Host–Microbe Interactions'. *Nature Reviews Genetics* 22 (6): 361–78. https://doi.org/10.1038/s41576-021-00326-y.

Wijayawardene, Nn, Kd Hyde, Dq Dai, M Sánchez-García, Bt Goto, Rk Saxena, M Erdoğdu, et al. 2022. 'Outline of Fungi and Fungus-like Taxa – 2021'. *Mycosphere* 13 (1): 53–453. https://doi.org/10.5943/mycosphere/13/1/2.

Woese, Carl R., and George E. Fox. 1977. 'Phylogenetic Structure of the Prokaryotic Domain: The Primary Kingdoms'. *Proceedings of the National Academy of Sciences* 74 (11): 5088–90. https://doi.org/10.1073/pnas.74.11.5088.

Woese, Carl R., O Kandler, and M L Wheelis. 1990. 'Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya.' *Proceedings of the National Academy of Sciences of the United States of America* 87 (12): 4576–79.

Wong, J.T., J. De Bruyn, B. Bagnol, H. Grieve, M. Li, R. Pym, and R.G. Alders. 2017. 'Small-Scale Poultry and Food Security in Resource-Poor Settings: A Review'. *Global Food Security* 15 (December):43–52. https://doi.org/10.1016/j.gfs.2017.04.003.

Wood, Derrick E., Jennifer Lu, and Ben Langmead. 2019. 'Improved Metagenomic Analysis with Kraken 2'. *Genome Biology* 20 (1): 257. https://doi.org/10.1186/s13059-019-1891-0.

Wright, Robyn J., André M. Comeau, and Morgan G. I. Langille. 2023. 'From Defaults to Databases: Parameter and Database Choice Dramatically Impact the Performance of Metagenomic Taxonomic Classification Tools'. *Microbial Genomics* 9 (3): 000949. https://doi.org/10.1099/mgen.0.000949.

Wu, Yu-Wei, Blake A. Simmons, and Steven W. Singer. 2016. 'MaxBin 2.0: An Automated Binning Algorithm to Recover Genomes from Multiple Metagenomic Datasets'. *Bioinformatics* 32 (4): 605–7. https://doi.org/10.1093/bioinformatics/btv638.

Xiao, Liang, Jordi Estellé, Pia Kiilerich, Yuliaxis Ramayo-Caldas, Zhongkui Xia, Qiang Feng, Suisha Liang, et al. 2016. 'A Reference Gene Catalogue of the Pig Gut Microbiome'. *Nature Microbiology* 1 (12): 1–6. https://doi.org/10.1038/nmicrobiol.2016.161.

Xie, Fei, Wei Jin, Huazhe Si, Yuan Yuan, Ye Tao, Junhua Liu, Xiaoxu Wang, et al. 2021. 'An Integrated Gene Catalog and over 10,000 Metagenome-Assembled Genomes from the Gastrointestinal

Microbiome of Ruminants'. *Microbiome* 9 (1): 137. https://doi.org/10.1186/s40168-021-01078-x.

Xue, Ming-Yuan, Hui-Zeng Sun, Xue-Hui Wu, Jian-Xin Liu, and Le Luo Guan. 2020. 'Multi-Omics Reveals That the Rumen Microbiome and Its Metabolome Together with the Host Metabolome Contribute to Individualized Dairy Cow Performance'. *Microbiome* 8 (1): 64. https://doi.org/10.1186/s40168-020-00819-8.

Yan, Wei, Congjiao Sun, Jingwei Yuan, and Ning Yang. 2017. 'Gut Metagenomic Analysis Reveals Prominent Roles of Lactobacillus and Cecal Microbiota in Chicken Feed Efficiency'. *Scientific Reports* 7 (1): 45308. https://doi.org/10.1038/srep45308.

Yang, Chao, Debajyoti Chowdhury, Zhenmiao Zhang, William K. Cheung, Aiping Lu, Zhaoxiang Bian, and Lu Zhang. 2021. 'A Review of Computational Tools for Generating Metagenome-Assembled Genomes from Metagenomic Sequencing Data'. *Computational and Structural Biotechnology Journal* 19 (January):6301–14. https://doi.org/10.1016/j.csbj.2021.11.028.

Yang, Jintao, Cuihong Tong, Danyu Xiao, Longfei Xie, Ruonan Zhao, Zhipeng Huo, Ziyun Tang, Jie Hao, Zhenling Zeng, and Wenguang Xiong. 2022. 'Metagenomic Insights into Chicken Gut Antibiotic Resistomes and Microbiomes'. *Microbiology Spectrum* 10 (2): e01907-21. https://doi.org/10.1128/spectrum.01907-21.

Yilmaz, Pelin, Laura Wegener Parfrey, Pablo Yarza, Jan Gerken, Elmar Pruesse, Christian Quast, Timmy Schweer, Jörg Peplies, Wolfgang Ludwig, and Frank Oliver Glöckner. 2014. 'The SILVA and "All-Species Living Tree Project (LTP)" Taxonomic Frameworks'. *Nucleic Acids Research* 42 (D1): D643–48. https://doi.org/10.1093/nar/gkt1209.

York, Ashley. 2019. 'Your Microbiome Is What You Eat'. *Nature Reviews Microbiology* 17 (12): 721–721. https://doi.org/10.1038/s41579-019-0287-1.

Zerjal, Tatiana, Sonja Härtle, David Gourichon, Vanaïque Guillory, Nicolas Bruneau, Denis Laloë, Marie-Hélène Pinard-van der Laan, Sascha Trapp, Bertrand Bed'hom, and Pascale Quéré. 2021. 'Assessment of Trade-Offs between Feed Efficiency, Growth-Related Traits, and Immune Activity in Experimental Lines of Layer Chickens'. *Genetics Selection Evolution* 53 (1): 44. https://doi.org/10.1186/s12711-021-00636-z.

Zhang, Han, Tanner Yohe, Le Huang, Sarah Entwistle, Peizhi Wu, Zhenglu Yang, Peter K. Busk, Ying Xu, and Yanbin Yin. 2018. 'dbCAN2: A Meta Server for Automated Carbohydrate-Active Enzyme Annotation'. *Nucleic Acids Research* 46 (W1): W95–101. https://doi.org/10.1093/nar/gky418.

Zhang, J., K. Kobert, T. Flouri, and A. Stamatakis. 2014. 'PEAR: A Fast and Accurate Illumina Paired-End reAd mergeR'. *Bioinformatics* 30 (5): 614–20. https://doi.org/10.1093/bioinformatics/btt593.

Zhang, Yan, Fan Jiang, Boyuan Yang, Sen Wang, Hengchao Wang, Anqi Wang, Dong Xu, and Wei Fan. 2022. 'Improved Microbial Genomes and Gene Catalog of the Chicken Gut from Metagenomic Sequencing of High-Fidelity Long Reads'. *GigaScience* 11 (November):giac116. https://doi.org/10.1093/gigascience/giac116.

Zhang, Yancong, Kelsey N Thompson, Curtis Huttenhower, and Eric A Franzosa. 2021. 'Statistical Approaches for Differential Expression Analysis in Metatranscriptomics'. *Bioinformatics* 37 (Supplement_1): i34–41. https://doi.org/10.1093/bioinformatics/btab327.

Zheng, Jinfang, Qiwei Ge, Yuchen Yan, Xinpeng Zhang, Le Huang, and Yanbin Yin. 2023. 'dbCAN3: Automated Carbohydrate-Active Enzyme and Substrate Annotation'. *Nucleic Acids Research* 51 (W1): W115–21. https://doi.org/10.1093/nar/gkad328.

Zheng, Jinfang, Boyang Hu, Xinpeng Zhang, Qiwei Ge, Yuchen Yan, Jerry Akresi, Ved Piyush, Le Huang, and Yanbin Yin. 2023. 'dbCAN-Seq Update: CAZyme Gene Clusters and Substrates in Microbiomes'. *Nucleic Acids Research* 51 (D1): D557–63. https://doi.org/10.1093/nar/gkac1068.