Vers la maîtrise des communautés microbiennes lignocellulolytiques : impact de la source d'inoculum et du prétraitement du substrat sur le fonctionnement des communautés

Lucas Auer

Université Fédérale
Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National des Sciences Appliquées de Toulouse (INSA de Toulouse)

Présentée et soutenue par :
AUER Lucas
le lundi 3 octobre 2016

Titre :

Vers la maîtrise des communautés microbiennes lignocellulolytiques : impact de la source d'inoculum et du prétraitement du substrat sur le fonctionnement des communautés

---

École doctorale et discipline ou spécialité :
ED SEVAB : Ingénieries microbienne et enzymatique

Unité de recherche :
Laboratoire d'Ingénierie des Systèmes Biologiques et des Procédés (LISBP)

Directeur/trice(s) de Thèse :

HERNANDEZ-RAQUET Guillermina
O'DONOHUE Michael

Jury :
PEYRET Pierre, Professeur des Universités, Clermont-Ferrand, rapporteur
SIMONET Pascal, Directeur de recherche, CNRS Lyon - rapporteur
COMBES Sylvie, Directeur de recherche, INRA Toulouse - examinateur
DABERT Patrick, Directeur de recherche, IRSTEA Rennes, examinateur
TERRAT Sébastien, Maître de conférence, INRA Dijon - examinateur

Résumé

La lignocellulose, composant majoritaire des parois végétales, peut être envisagée comme une immense source de carbone renouvelable. Cependant, elle est encore largement sous-exploitée à l'heure actuelle, notamment du fait des limitations et du coût des approches classiques de libération des sucres qui la composent, et de la difficulté à la dégrader. Pourtant, dans la nature, les substrats lignocellulosiques sont à la base de nombreuses chaînes alimentaires où des micro-organismes ont donc le potentiel de les transformer intégralement. Dans cette thèse, nous nous sommes intéressés à deux sources naturelles de communautés microbiennes lignocellulolytiques (rumen bovin et intestin de termite), à leur potentiel de dégradation de lignocellulose en bioréacteur, et surtout, à l'évolution et aux dynamiques de leurs communautés microbiennes mises en cultures. La mise au point d'une méthode de traitement de données de séquençage capable de gérer des jeux de données de grande taille a permis d'étudier ces dynamiques par séquençage Illumina de la région V3-V4 des ADNr 16S et ARNr 16S. L'étude de l'effet de modification du substrat par prétraitement sur les communautés et leur dynamique d'une part, et l'étude de l'effet de la source d'inoculum ont permis d'apporter de nouveaux éléments de compréhension sur le fonctionnement des écosystèmes microbien lignocellulolytiques en bioréacteur, et notamment sur de probables relations de complémentarité fonctionnelle entre différents phylum bactériens.

Mots clés : lignocellulose, paille, bioréacteur, fermentation, AGV, diversité, communautés bactériennes, ARN 16S, amplicon, metabarcoding, singletons, rumen, termite.

Lignocellulose is the main component of plant cell wall but can also be seen as a wide renewable carbon source. Nevertheless, it is still currently underexploited due to the costs and limitations of classical conversion approaches and to its recalcitrance to degradation. However, in Nature, lignocellulosic substrates are the basis of several food chains where microorganisms have the potential to completely transform them. In this thesis, we were interested in two natural sources of lignocellulolytic microbial communities (bovine rumen and termite gut), in their lignocellulose degradation abilities and above all, in the evolution and dynamics of their microbial communities once in bioreactors. The development of a sequencing data processing method able to deal with large datasets allowed studying these dynamics thanks to Illumina sequencing of the V3-V4 region of 16S rDNA and 16S rRNA. The study of the effect of substrate modification using pretreatments on communities and their dynamics, and of the effect of inoculum sources gave new insights into lignocellulolytic microbial ecosystems functionment in bioreactors, and especially into likely functional interplays between different bacterial phyla.

Key words : lignocellulose, wheat straw, bioreactors, fermentation, VFA, diversity, bacterial communities, 16S RNA, amplicon, metabarcoding, singletons, rumen, termite gut.

Productions scientifiques

Publications scientifiques en premier auteur

Auer, L. & Lazuka, A., Bozonnet, S., Morgavi, D.P., O'Donohue, M., and Hernandez-Raquet, G. (2015). Efficient anaerobic transformation of raw wheat straw by a robust cow rumen-derived microbial consortium. Bioresour. Technol. 196, 241–249. **Published**

Auer, L., Mariadassou, M., O'Donohue, M., Klopp, C., and Hernandez-Raquet, G. Singleton read filtering of large 16S rRNA Illumina datasets enable fast and accurate microbial community description. Molecular Ecology Resources. **Published**

Auer, L. & Lazuka, O'Donohue, M., and Hernandez-Raquet, G. Diversity of metabolically active bacteria associated with lignocellulose degradation by a cow rumen-derived consortium: dynamic responses to substrate pretreatment. **In preparation**

Auer, L., Lazuka, A., Sillam-Dusses, D., Miambi, E., Dumas, C., O'Donohue, M., and Hernandez-Raquet, G. Exploring the potential of termite gut microbiota as biocatalyst for lignocellulose bioconversion. **In preparation**

Auer, L., Lazuka, O'Donohue, M., and Hernandez-Raquet, G. Community dynamics of a lignocellulolytic bacterial consortium derived from termite-gut microbiome. **In preparation**

Auer, L. & Escudié, F., Bernard, M., Cauquil, L., Vidal, K., Maman, S., Mariadassou, M., Hernandez-Raquet, G., and Pascal, G. Presenting FROGS, a fast and accurate tool for metabarcoding analyses of large sequencing datasets. **Submitted**

Publications scientifiques (contributions)

Peyre-Lavigne, M., Bertron, A., Auer, L., Hernandez-Raquet, G., Foussard, JN., Escadeillas, G., Cockx, A., and Paul, E. (2015). Innovative approach to reproduce the biodeterioration of industrial cementitious products in a sewer environment. Part I: Test design. Cement and Concrete Research

Peyre-Lavigne, M., Bertron, A., Auer, L., Hernandez-Raquet, G., Foussard, JN., Escadeillas, G., Cockx, A., and Paul, E. (2016). Innovative approach to reproduce the biodeterioration of industrial cementitious products in a sewer environment. Part II: Validation on CAC and BFSC linings. Cement and Concrete Research

Abot, A., Arnal, G., Auer, L., Lazuka, A., Labourdette, D., Lamarre, S., Trouilh, L., Laville, E., Lombard, V., Potocki-Veronese, G., Henrissat, B., O'Donohue, M., Hernandez-Raquet, G., Dumon, C. and Leberre, V. (2016). *CAZyChip : dynamic assessment of exploration of glycoside hydrolases in microbial ecosystems.* BMC Genomics

Conférences

Auer, L., Lazuka, A., Paisse, S., O'Donohue, M., and Hernandez-Raquet, G. *Valorisation de la lignocellulose : analyse fonctionnelle de consortia microbien enrichis*. Génomique Environnementale, 4-6 novembre 2013, Rennes. **Poster**.

Auer, L., Lazuka, A., Abadie, M., O'Donohue, M., and Hernandez-Raquet, G. *Lignocellulose degradation by engineered microbial consortia from rumen and termite gut: correlating enzymatic profiles and functional microbial diversity*. International Society for Microbial Ecology symposia, 24-29 August 2014, Seoul. **Oral presentation**.

Auer, L. & Escudié, F., Bernard, M., Cauquil, L., Vidal, K., Maman, S., Mariadassou, M., Hernandez-Raquet, G., and Pascal, G. *FROGS, Find Rapidly OTU with Galaxy Solution.* International Pathobiome congress, 24-26 June 2015, Paris. **Oral presentation**.

Auer, L., Cauquil, L., Chaillou, S., Delbes, C., Dugat-Bony, E., Falentin, H., Hernandez-Raquet, G., Mariadassou, M., Nicolas, A., Pascal, G., Rifa, E., Schbath, S., Abraham, AL., Terrat, S. *How to design an efficient and robust pipeline for 16S rRNA-gene sequence analysis to improve our understanding on microbial communities?* Journées ouvertes de bioinformatique et mathématiques, 6-9 juillet 2015. **Poster**.

Auer, L., Lazuka, A.,Oelker, G., Jehmlich, N., O'Donohue, M., and Hernandez-Raquet, G. *Analyses méta-omiques de communautés bactériennes issues du microbiote intestinal de termite impliquées dans la production de carboxylates à partir de lignocellulose.*Génomique environnementale, 26-28 octobre 2015, Montpellier. **Oral presentation**.

Auer, L., Lazuka, A., Flajollet, E., O'Donohue, M., Jehmlich, N., von Bergen, M., and Hernandez-Raquet, G. *Can dry chemo-chemical pretreatment increase lignocellulose digestibility by microbial consortia? Impact on diversity and CAZyme profiles.* International Society for Microbial Ecology symposia, 21-26 August 2016, Montreal. **Poster**. Granted by an ISME Committee Travel Award.

Sommaire

Productions scientifiques 5
Publications scientifiques en premier auteur 8
Publications scientifiques (contributions) 8
Conférences 9

REMERCIEMENTS 16

INTRODUCTION 21

CHAPITRE I : SYNTHESE BIBLIOGRAPHIQUE 29
I. La lignocellulose et sa dégradation 29
I.1. Constituants de la lignocellulose 29
I.1.1 Fraction cellulosique 30
I.1.2 Fraction hémi-cellulosique 31
I.1.3 Lignine 32
I.1.4 Structure 33
I.2. Procédés de valorisation actuels 34
I.2.1 Compostage et méthanisation 34
I.2.2 Plateforme des sucres 35
I.2.3 Plateforme des carboxylates 36
I.3. Prétraitements 38
I.3.1 Efficacité des prétraitements de la biomasse 38
I.3.2 Méthodes de prétraitement 39
I.3.3 Méthodes physiques de prétraitement 39
I.3.4 Méthodes chimiques de prétraitement 40
I.3.5 Méthodes physico-chimiques 41
I.3.6 Autres méthodes 42
II. Dégradation naturelle de la lignocellulose 44
II.1. Digestion de la biomasse chez les mammifères ruminants 44
II.2. Insectes 46
II.2.1 Les Termites 46
II.2.2 Termites inférieurs 47
II.2.3 Termites supérieurs 49
II.3. Études de la diversité associée à la dégradation de la lignocellulose 50
II.3.1 Exemple en système naturel : le rumen 50
II.3.2 Études en systèmes artificiels : bioréacteurs 54
.II.3.2.1 Enrichissements 54
.II.3.2.2 Évolution de la diversité au cours de la dégradation 59
.II.3.2.3 Effet des prétraitements 60
III. Méthodes et outils d'étude des communautés bactériennes 63
III.1. Gènes utilisés en écologie moléculaire 64
III.1.1 ARNr 16S 64
III.1.2 Autres marqueurs 66
III.1.3 Banques de données 68
III.2. Techniques de suivi 69
III.2.1 Les techniques sans séquençage 69
III.2.2 Techniques de séquençage 71
.III.2.2.1 Séquençage Sanger 71
.III.2.2.2 Le séquençage nouvelle génération (Next Genation Sequencing) 71

.III.2.2.3 Roche 454 system 72
.III.2.2.4 Illumina GA/HiSeq System 72
.III.2.2.5 Autres technologies 76
III.3. Analyse des données amplicon générés par NGS 78
III.3.1 Étapes clés et outils dédiés 78
.III.3.1.1 Nettoyage des données 78
.III.3.1.2 Déchimérisation 79
.III.3.1.3 Clustering 81
.III.3.1.4 Assignation taxonomique 86
III.3.2 Outils d'analyse de séquences généraux 87
.III.3.2.1 Mothur 87
.III.3.2.2 QIIME 88
III.4. Outils statistiques pour l'écologie moléculaire 88
III.4.1 Indices (diversité α) 88
III.4.2 Distances (diversité β) 90
III.4.3 Méthodes de projection et visualisation 91

CHAPITRE II : PROCEDURES EXPERIMENTALES 95
I. Origine des populations bactériennes 95
I.1. Rumen initial 95
I.2. Termites 95
I.2.1 Choix des espèces 95
I.2.2 Récolte et conditionnement 96
II. Dispositifs expérimentaux de culture en bioréacteur 97
II.1. Bioréacteurs du screening termites 97
II.1.1 Dispositif expérimental 97
II.1.2 Paramètres suivis 97
II.2. Bioréacteurs d'enrichissement rumen et termite 98
II.2.1 Dispositif expérimental 98
II.2.2 Paramètres suivis 98
II.2.3 Bioréacteurs paille stérile / non stérile (termites) 99
II.3. Bioréacteurs de cinétiques de caractérisation 99
II.3.1 Dispositif expérimental 99
II.3.2 Paramètres suivis 99
II.3.3 Prétraitements 100
II.3.4 Plan d'expériences 101
III. Analyses macrocinétiques 102
III.1. Mesure des matières volatiles résiduelles 102
III.2. Dosage des acides gras volatils produits 102
III.3. Dosage des sucres résiduels 102
III.4. Mesure des activités enzymatiques 103
IV. Outils moléculaires 103
IV.1. Méthode de co-extraction ADN/ARN 103
IV.2. Analyses par amplification de l'ADN 16S 104
IV.2.1 Dosage des copies 16S par qPCR 104
IV.2.2 Préparation des librairies MiSeq Illumina 104
.IV.2.2.1 PCR1 de préparation des librairies (amplification V3-V4) 104
.IV.2.2.2 PCR2 de préparation des librairies (ajout index et adaptateurs) 105
.IV.2.2.3 Lancement du séquençage sur MiSeq 105
V. Traitement des données de séquençage (pipeline IPS) 106
V.1. Formatage et création d'un fichier fasta unique 106
V.2. Nettoyage des données, pre-clustering, filtres de bruit et dé-chimérisation 106

V.3. Assignation taxonomique des séquences 107
V.4. Calcul des distances et clustering 108
V.5. Assignation taxonomique des OTUs et calcul de distance entre OTUs 108
V.6. Résultats, sorties graphiques et import R 108
VI. Analyse comparatives des dynamiques de communautés bactériennes 109
VI.1. Calcul des indices de diversité 109
VI.2. Statistiques exploratoires 109
VI.2.1 Analyses de projection de données (PCA et PLS) 109
VI.2.2 Méthodes de projections de distances 110
VI.3. Statistiques de test et classification supervisée 110
VI.3.1 ANOVA / PERMANOVA 110
VI.3.2 sPLS-DA 111

CHAPITRE III : VALIDATION DES METHODES DE TRAITEMENT DE DONNEES DE SEQUENCAGE 115
I. Introduction 115
II. Analyse de gros jeux de données 16S Illumina : impact d'un filtre de séquence singleton sur la description des communautés microbiennes. 116
II.1. Abstract 116
II.2. Introduction 117
II.3. Results 119
II.3.1 Simulated datasets 120
II.3.2 Mock communities 124
II.3.3 Real datasets 126
II.3.4 Stress datasets 127
II.4. Discussion 128
II.4.1 Chimeras 128
II.4.2 Contaminants 129
II.4.3 Effect on community reconstruction 130
II.4.4 Rare filter 131
II.5. Material and methods 131
II.5.1 Simulated data 131
II.5.2 Sequencing 132
II.5.3 Data processing 133
II.5.4 Diversity analyses 134
II.6. Acknowledgments 134
II.7. References 135
II.8. Supplementary data 137
III. Conclusion du chapitre 145

CHAPITRE IV : STABILISATION ET CARACTERISATION D'UN INOCULUM ISSU DE RUMEN BOVIN 149
I. Introduction 149
II. Efficient anaerobic transformation of raw wheat straw by a robust cow rumen-derived microbial consortium 150
II.1. Abstract 150
II.2. Introduction 151
II.3. Material and methods 153
II.3.1 Lignocellulosic substrate and inoculum 153
II.3.2 Anaerobic reactors 153
II.3.3 Chemical analyses 154
II.3.4 Enzyme activity assays 155

II.3.5 Analysis of microbial diversity 156
II.4. Results and discussion 157
II.4.1 Enrichment: macro-kinetic and diversity analysis 157
II.4.2 Characterization of wheat straw degradation by consortium RWS 162
.II.4.2.1 Wheat straw degradadation and VFA production by RWS 162
.II.4.2.2 Effects of RWS activity on the physicochemical composition of wheat straw 164
.II.4.2.3 Enzymatic activity profiles along wheat straw degradation 167
II.5. Conclusion 171
II.6. Acknowledgements 171
II.7. References 172
III. Conclusion du chapitre 175

CHAPITRE V : ETUDE DE L'IMPACT DU PRETRAITEMENT DU SUBSTRAT SUR LA DYNAMIQUE DES COMMUNAUTES BACTERIENNES AU COURS DE LA DEGRADATION 179
I. Introduction du chapitre 179
II. Effect of lignocellulosic substrate pretreatment on lignocellulolytic community dynamics 181
II.1. Abstract 181
II.2. Introduction 182
II.3. Methods 184
II.3.1 Lignocellulosic pretreated substrates 184
II.3.2 Lignocellulose degradation reactors 184
II.3.3 Nucleic acids isolation 185
II.3.4 DNA and cDNA amplification and sequencing 185
II.3.5 Data processing 186
II.3.6 Diversity analysis 187
II.4. Results 187
II.4.1 Degradation of pretreated wheat straw substrates by RWS 187
II.4.2 Microbial community structure on different pretreated substrates 188
II.4.3 Community dynamics during LC degradation 192
II.5. Discussion 195
II.5.1 Correlation between community dynamics and fermentation parameters 195
II.5.2 Cyclic 16S rRNA profiles 196
II.5.3 OTUs specific of a pretreatment 197
II.5.4 OTUs related to specific LC degradation phase 198
II.6. Conclusion 199
II.7. Acknowledgments 199
II.8. References 200
II.9. Supplementary data 203
III. Conclusion du chapitre 206

CHAPITRE VI : EXPLORATION DU POTENTIEL DE DEGRADATION DE FLORES INTESTINALES DE TERMITES 209
I. Introduction 209
II. Exploration du potentiel de microbiote intestinal de termite comme biocatalyseur de la bioconversion de lignocellulose 210
II.1. Abstract 210
II.2. Introduction 211
II.3. Results 213
II.3.1 Reactor performances: degradation and products 213
II.3.2 Diversity analysis changes of termites gut inocula after lignocellulose degradation 216
II.3.3 Gut flora composition 216

II.3.4 Diversity changes after incubation in lignocellulose reactors 218
II.4. Discussion 221
II.5. Conclusion 226
II.6. Experimental procedures 226
II.6.1 Lignocellulose substrate and termite gut inocula 226
II.6.2 Anaerobic reactors 227
II.6.3 Chemical analysis 228
II.6.4 Enzyme activity assay 228
II.6.5 16 rRNA gene copy number and diversity analysis 229
II.7. Acknowledgments 231
II.8. References 232
II.9. Supplementary data 237
III. Conclusion du chapitre 239

CHAPITRE VII : ANALYSE DE L'EVOLUTION DE LA DIVERSITE AU COURS DE L'ENRICHISSEMENT A PARTIR DE MICROBIOME INTESTINAL DE TERMITES 243
IV. Introduction 243
V. Dynamique des communautés d'un consortium bactérien dérivé du microbiote intestinal de termite. 244
V.1. Introduction 244
V.2. Methods 246
V.2.1 Enrichment of termite-gut community in lignocellulose transformation reactors 246
V.2.2 Diversity analyses 247
V.2.3 Data processing 248
V.3. Results 249
V.3.1 Termite-gut microbiome enrichment 249
V.3.2 Diversity dynamics throughout lignocellulose degradation by TWS 252
V.3.3 Comparing the diversity dynamics of termite-derived and rumen-derived consortia along lignocellulose transformation. 254
V.3.4 Identification of OTUs common and specific to TWS and RWS 256
V.4. Discussion 260
V.5. Conclusion 263
V.6. Acknowledgments 263
V.7. References 264
V.8. Supplementary data 266
VI. Conclusion du chapitre 270

CONCLUSION ET PERSPECTIVES 273

REFERENCES BIBLIOGRAPHIQUES 281

REMERCIEMENTS

« Un jour, j'irai vivre en Théorie. Parce qu'en Théorie, tout se passe bien. »

Le chemin qui y mène serait long et difficile tout seul, sans personne pour nous orienter ou nous accompagner, et sans ange gardien, bonne étoile ou chance selon les appellations, pour causer les bonnes rencontres aux bons moments. Impossible en tout cas d'en arriver là sans être redevable à quiconque et avoir à remercier qui que ce soit. Dans mon cas, mon ange gardien et moi-même tenons à remercier en premier lieu mes parents et grands-parents, qui ont toujours voulu et sacrifié pour que leurs enfants puissent faire les études qu'ils souhaitaient, dans la direction qu'ils souhaitaient, pour faire le métier qu'ils souhaitaient. Mais également mes amis de lycée, puis camarades de promotion en prépa au Lycée Henri Poincaré et à l'ENS de Cachan, auprès de qui s'est peu à peu construit le parcours qui m'a amené jusqu'ici. Sans oublier les professeurs géniaux rencontrés au cours de mes études, ceux qui nous font nous passionner pour leur matière et nous donnent envie d'en découvrir toujours plus, et on fait naître cette envie de recherche.

Je tiens évidemment ensuite à remercier les nombreuses personnes qui ont contribué à leur niveau à l'aboutissement de mon travail de thèse. Notamment ma directrice, Guillermina, pour son implication et son accompagnement quasiment quotidien. Tout l'inverse d'un directeur absent, elle m'a donné la chance de travailler presque sans limites financières, de suivre autant de formations que je souhaitais, de m'investir dans des groupes de travail et de me faire participer à de nombreux congrès. Elle a su me donner sa confiance, et nous avons toujours réussi à discuter science, enaccord ou en désaccord, de la thèse ou de thématiques complètement différentes. Je remercie aussi Mike pour le temps qu'il a pu nous consacrer malgré ses responsabilités toujours plus grandes, et pour son aide dans l'amélioration d'articles, son regard critique et ses remarques pertinentes.

Je remercie Adèle bien sûr, sans qui mon travail de thèse n'aurait pas été possible. Binôme dès mon arrivée, mais aussi soutien moral réciproque au quotidien (notamment pour les repiquages nocturnes du début de thèse). Notre complémentarité et celle de nos thèses ont été un vrai atout pour ces années de doctorat. Parce que les stagiaires sont un des piliers de notre monde de la recherche, merci à Julie, Maïder, Cécile, Gunnar, Anaïs, pour votre aide, directe ou indirecte et pour les moments passés ensembles. Merci à tout le reste de l'équipe, membres permanents ou précaires de la recherche, pour leur participation de près ou de loin, directement ou indirectement, à la réussite professionnelle et personnelle de mes années de thèse.

Merci ensuite à tous les collaborateurs avec qui j'ai eu l'occasion de travailler ou d'échanger, notamment à la plateforme de séquençage ou à l'IRD, membres du PEPI ou de la plateforme bioinfo. Un gros merci à l'équipe de FROGS pour son travail et sa bonne humeur, et tout spécialement à Géraldine et Laurent pour nos discussions passionnées sur le clustering, la formation des chimères, les filtres ou les données simulées dès les débuts de ma thèse. C'était un plaisir de découvrir la métagénomique 16S à vos côtés, et de voir le têtard se transformer en vraie grenouille. J'espère pouvoir continuer à m'impliquer dans le projet pour la suite.

Parce qu'un doctorat est également l'occasion de nouer des relations bien plus que professionnelles, je tiens aussi à remercier Mourad pour son amitié sincère, pour le partage entre nos deux cultures, que ce soit personnelles ou professionnelles, et pour sa compagnie pendant toutes nos années comme voisins de bureau, et même après. Je remercie Émeline également, pour son affection, sa bonne humeur, sa compréhension et son soutien pendant la dernière année de thèse, la plus longue et la plus difficile.

Enfin, pour boucler la boucle, je souhaite encore remercier ma famille, et notamment les jeunes parents de mes neveux et nièces, qui m'ont permis de garder à l'esprit que la vraie vie des vrais gens continuait d'exister quelque part, pas très loin là au dehors. Merci à eux, et à tous les êtres proches qui ont dû supporter à des degrés divers les effets de l'investissement dans l'aventure qu'est une thèse. J'espère pouvoir toujours être aussi compréhensif et disponible pour vous et que vous avez pu l'être pour moi.

INTRODUCTION

INTRODUCTION

Les microorganismes, dont l'existence a longtemps été ignorée, ont pourtant été les premiers à occuper la planète et ont contribué à façonner notre environnement. Leur évolution leur a permis de coloniser tous les écosystèmes, même les plus extrêmes. Dotés d'une grande plasticité génétique et génomique et d'une exceptionnelle diversité métabolique, ils se sont adaptés à tous les bouleversements climatiques et géologiques survenus sur Terre dans les trois derniers milliards d'années. Avec l'apparition d'organismes macroscopiques, leur rôle dans les écosystèmes n'en est pas devenu moins important. Au contraire, les microorganismes jouent un rôle primordial à la base de nombreuses chaînes alimentaires, et occupent également une position clé dans tous les cycles du carbone ou de l'azote en tant que recycleur. Enfin, ils colonisent également les autres êtres vivants et sont alors impliqués dans d'innombrables interactions, potentiellement bénéfiques (on parle alors de mutualisme et symbiose). Dans ce contexte, l'adaptabilité et les grandes potentialités métaboliques des microorganismes sont mis au service de l'organisme hôte qui ne les possède pas. Ainsi, chez les animaux la flore intestinale microbienne aide fortement à la digestion. Dans le cas des herbivores, ceux-ci sont même incapables de digérer les végétaux sans l'aide de leur microbiote intestinal, alors qu'ils constituent pourtant parfois leur seule alimentation. Les interactions avec des microorganismes ne se limitent pas aux flores digestives. Elles peuvent concerner des plantes (les Légumineuses sont capables de fixer l'azote atmosphérique grâce à une symbiose avec des bactéries), permettre le développement des coraux (symbiose entre un cnidaire et une algue unicellulaire) ou encore former un organisme composite (lichen formé par la symbiose entre un champignon et une algue ou une cyanobactérie). Ces interactions avec un microorganisme peuvent également se traduire par des effets néfastes, pouvant conduire à des pathologies, la mort de l'hôte, ou des effets plus surprenants, comme des modifications de comportement induites par Toxoplasma, un parasite des chats. L'étude des interactions impliquant un microorganisme, appelée écologie microbienne, est donc un champ de recherche extrêmement vaste et contemporain, ceux-ci n'ayant été découverts et étudiés qu'avec la science moderne.

Pourtant, si l'implication des microorganismes dans des phénomènes visibles n'a été découverte qu'au XIX^{e} siècle, leurs potentialités sont en fait utilisées depuis bien plus

longtemps. La fermentation de fruits, végétaux ou lait en vin, alcools divers ou fromage est maîtrisée dès l'Antiquité par diverses civilisations, sans avoir connaissance de la microbiologie sous-jacente. Plus récemment, les procédés d'épuration de l'eau ont été développés, s'appuyant sur des communautés microbiennes sans aucune connaissance de leur fonctionnement. C'est là un avantage de l'utilisation de ces communautés : leur adaptabilité est tellement grande qu'il est possible de sélectionner une population réalisant une fonction sans même la connaître. Les progrès en microbiologie, puis en génétique, ont permis l'essor d'un biotechnologie d'un nouveau genre, utilisant le potentiel des microorganismes mais s'appuyant sur des souches bactériennes isolées, là où les processus naturels font généralement intervenir des communautés complexes. En biotechnologie comme en agriculture, l'utilisation de souches ou variétés uniques a l'avantage de permettre un contrôle plus fort, une optimisation plus facile, et une production plus régulière. Par contre, une telle approche diminue aussi l'adaptabilité du système, ce qui l'expose à des risques en cas de perturbations ou diminue ses performances face à des conditions environnementales fluctuantes.

Les biotechnologies ont de nombreux domaines d'application, allant du domaine médical (synthèse de l'insuline pour le traitement des diabétiques) au développement d'OGM végétaux (plants de café sans caféine). Plus récemment, elles ont permis le développement des filières de biocarburants. En effet, l'accroissement constant de la demande mondiale en énergie, couplée à l'épuisement des ressources, à la fluctuation des prix du pétrole, et au développement d'une certaine prise de conscience écologique a conduit à rechercher des sources d'énergie alternatives. Dans ce contexte, la transformation de produits végétaux, alors vus comme sources de carbone et d'énergie renouvelables par biotechnologie a conduit à la production de biocarburants de première génération. Ceux-ci consistent en une simple transformation de sucre en bioéthanol. Le sucre étant alors produit par des surfaces agricoles dédiées, la production de biocarburants de première génération est polémique puisqu'elle implique une compétition entre production alimentaire et production d'énergie. Les biocarburants de deuxième génération, eux, sont issus de la transformation de résidus agricoles ou de parties non alimentaires de plantes agricoles. Cependant, ces substrats présentent des rendements de conversion bien plus faibles, car ils sont plus difficiles à transformer que le sucre. Bien qu'ils représentent une ressource non valorisée disponible en quantités presque illimitées, ces substrats sont constitués principalement de lignocellulose, le biopolymère le plus abondant sur Terre. La lignocellulose est une structure physico-chimique

complexe, impliquant différentes molécules (cellulose, hémicelluloses et lignines) liées entre elles, conférant à l'ensemble une résistance élevée aux attaques biotiques et abiotiques. Par sa variabilité de molécules, de liaisons et de structure, sa déconstruction requiert un grand nombre d'activités enzymatiques différentes. Les approches de biotechnologie classique (« plateforme des sucres »), qui font intervenir des traitements par des enzymes pour libérer les sucres constituant le substrat, atteignent difficilement des rendements élevés. De nombreuses recherches ont donc pour but d'améliorer la transformation de biomasse lignocellulosique, en utilisant différentes approches. L'utilisation de prétraitements permet d'améliorer l'accessibilité des différents composants aux enzymes, d'altérer la structure cristalline de la cellulose ou l'intégrité de la lignine. La sélection de nouvelles enzymes permet d'augmenter la diversité dans les cocktails enzymatiques et de les enrichir avec des enzymes de plus en plus actives. S'ils permettent d'augmenter les rendements de transformation, les coûts élevés du prétraitement et de l'utilisation massive d'enzymes, le tout en conditions stériles, deviennent alors les principaux verrous pour une bio-raffinerie de lignocellulose économiquement rentable.

La transformation de la lignocellulose est donc un processus difficile. Pourtant, les végétaux sont à la base des pyramides alimentaires de la majorité des écosystèmes terrestres, et la lignocellulose qu'ils produisent y est donc dégradée à un niveau ou un autre. Cette dégradation peut prendre de nombreuses formes selon les environnements : méthanisation (marais), acétogénèse (flore intestinale d'animaux), humification ou minéralisation (sols et sédiments). Ces différents environnements ont déjà été étudiés à la recherche d'enzymes plus performantes. Il est également possible d'envisager une utilisation directe de communautés microbiennes complexes dans des procédés de transformation. Les systèmes présentant les meilleurs rendements de conversion sont aérobies et aboutissent à la transformation du substrat en CO_2, ce qui a industriellement peu d'intérêt. Mais les systèmes anaérobies aboutissent à la production de molécules carbonées à courtes chaînes (acides gras volatils ou encore carboxylates), qui peuvent avoir un intérêt en industrie chimique, et peuvent également être orientées vers la production de biocarburants. L'utilisation directe de communautés microbiennes pour transformer la lignocellulose en conditions anaérobies (ou « plateforme carboxylate ») est donc une voie alternative à la plateforme des sucres. De nombreux travaux y sont déjà consacrés, s'attachant notamment à la sélection, la stabilisation et la caractérisation de communautés microbiennes performantes en culture, mais aussi à l'amélioration de leurs capacités de dégradation à l'aide de prétraitements du substrat.

Cependant, les outils de la microbiologie classique sont difficiles à appliquer à des communautés microbiennes complexes, et la fermentation avec un substrat solide pose de nombreuses contraintes techniques. La compréhension du fonctionnement des communautés microbiennes utilisées en plateforme carboxylate est donc encore très limitée. Par ailleurs, peu d'étude se sont encore intéressées à l'inoculation de fermenteurs par des microbiotes intestinaux, et aucune pour la production de carboxylates. Pourtant, ces écosystèmes sont parmi les plus performants du monde vivant et ont donc un potentiel de dégradation élevé. En permettant l'identification de similitudes et différences entre les communautés, l'utilisation et la comparaison d'inocula différents peut renforcer notre compréhension de leur fonctionnement.

En associant les outils d'écologie microbienne à une approche de génie des procédés, ces travaux de thèse s'intéressent à la caractérisation et à l'étude du fonctionnement des communautés microbiennes associées à la transformation d'un substrat lignocellulosique en conditions anaérobies. L'utilisation du microbiote ruminal bovin et intestinal de termite a permis d'étudier les performances de ces inoculums pour la dégradation en conditions de culture en fermenteurs, ce qui n'avait encore jamais été testé pour la plateforme carboxylates. Leur utilisation avec des substrats prétraités et le développement d'outils d'écologie microbienne dédiés à l'analyse des communautés, a également permis d'améliorer notre compréhension de ces écosystèmes par des approches comparatives. Ces travaux de thèse se concentrent principalement sur l'écologie microbienne de ces fermenteurs, sont complémentaires de ceux de la thèse d'Adèle Lazuka, orientés fermentation et procédés. Les expériences de fermentation (à l'exception de celles du chapitre VI, entièrement conduites par moi-même) ont été majoritairement menées par elle, avec des contributions plus mineures de ma part. Des informations plus riches sur les aspects procédés et caractérisation de la fermentation sont donc disponibles dans ses propres travaux de thèse, ce manuscrit s'attachant plus particulièrement à l'écologie microbienne de nos fermenteurs.

L'ensemble des données obtenues est présenté dans ce manuscrit sous forme de sept chapitres. Une synthèse bibliographique (chapitre I), divisée en trois parties, fait tout d'abord un bref état des connaissances acquises dans des travaux antérieurs. La première partie s'attache à décrire la lignocellulose et à présenter les procédés permettant sa valorisation. La deuxième partie permet une rapide présentation de la dégradation de la lignocellulose dans

quelques milieux naturels, avant de s'intéresser à l'état des connaissances sur les communautés microbiennes utilisées avec la plateforme carboxylate. Enfin, la troisième partie fait un état des lieux des technologies et méthodes d'analyse des communautés bactériennes. Le chapitre II présente les procédures expérimentales mises en place sur dans les chapitres suivants. Ceux-ci prennent la forme d'articles scientifiques qui ont été soumis (chapitre III), publiés (chapitre IV) ou sont en préparation (chapitre V, VI et VII). Une première étude (chapitre III) présente la validation d'une méthode d'analyse des données de séquençage permettant une diminution des capacités requises pour l'analyse de jeux de données de grande taille, sans altérer la qualité des résultats produits. Le développement de cette méthode a été nécessaire pour analyser simultanément plusieurs centaines d'échantillons dans les chapitres suivants. La deuxième étude (chapitre IV) présente les résultats de l'utilisation d'un inoculum issu de rumen bovin. Elle a permis la stabilisation d'une communauté lignocellulolytique, nommée RWS, utilisée dans les chapitres suivants. Le chapitre V présente les résultats de la culture de RWS sur des substrats prétraités, et de l'analyse de la dynamique des communautés pendant un cycle de dégradation, rendant possible l'étude de la réponse de la communauté à la modification du substrat par un prétraitement, ainsi que l'identification des acteurs majoritaires de la dégradation. Dans le chapitre VI, les capacités de dégradation de quatre microbiotes intestinaux de termites ont été testées en culture. Les communautés finales obtenues sont très différentes du microbiote initial, mais permettent tout de même d'obtenir une dégradation élevée de la lignocellulose. Enfin, le chapitre VII présente la stabilisation de la communauté la plus performante issue du microbiote intestinal de termite (obtenue dans le chapitre VI) pour obtenir la communauté TWS. Sa dynamique au cours d'un cycle de dégradation y est présentée et les données accumulées sur RWS et TWS sont comparées afin d'étudier l'effet de la source d'inoculum sur le fonctionnement des communautés stabilisées.

Ce travail s'inscrit dans les projets de recherche ProBioS et Insyme, et a été financé par l'Institut Carnot, la région Languedoc-Roussillon Midi-Pyrénées, et l'INRA. Il a été réalisé au sein du Laboratoire d'Ingénierie des Systèmes Biologiques et des Procédés (LISBP), dans l'équipe SYMBIOSE. Il a fait l'objet de collaborations avec les équipes de l'Institut pour la Recherche et le Développement (IRD) de Bondy et avec l'équipe NED (Nutrition et Ecosystèmes Digestifs) de l'INRA Toulouse.

CHAPITRE I

SYNTHESE BIBLIOGRAPHIQUE

CHAPITRE I : SYNTHESE BIBLIOGRAPHIQUE

I. La lignocellulose et sa dégradation

La lignocellulose est un macropolymère produit par les végétaux ; la lignocellulose est en effet le composant principal de la paroi cellulaire et peut représenter jusqu'à 90% du poids sec (Faraco, 2013). Elle est ainsi le biopolymère le plus abondant sur Terre. Sa principale fonction est d'assurer la résistance mécanique et physique permettant la turgescence des cellules végétales. Cette résistance est rendue possible par la composition et la structure de la lignocellulose, même si celles-ci varient en fonction des espèces végétales (Vassilev et al., 2012), des conditions locales, climatiques et écologiques. La résistance et la complexité de la lignocellulose la rendent également difficile à dégrader, ce qui présente des intérêts écologiques pour les plantes en termes de défenses et de maintien de leur intégrité, mais est un obstacle pour son utilisation comme source de carbone renouvelable (Chen et al., 2008). Pourtant, l'agriculture et les industries forestières produisent de larges quantités de déchets lignocellulosiques riches en carbone, principalement polysaccharides et composés aromatiques, qui peuvent être valorisés sous forme de biocarburants, bioplastiques et autres dérivés chimiques (synthons) d'intérêt industriel. La conversion de la lignocellulose en produits chimiques à haute valeur ajoutée est donc l'un des enjeux majeurs de notre siècle pour remplacer le flux des matières d'origine pétrolière par des ressources renouvelables et durables.

I.1. Constituants de la lignocellulose

La lignocellulose est principalement composée de trois polymères : cellulose, hémicellulose et lignine. Ceux-ci sont associés en une matrice hétérogène dont la cohésion est assurée à la fois par la résistance des différents composants et par des liaisons entre eux.

I.1.1 Fraction cellulosique

La cellulose est le constituant structural des cellules végétales. Son abondance varie entre 35 et 60% en fonction de l'espèce végétale ou de l'organe concerné, ce qui en fait le composant majoritaire de la lignocellulose. La cellulose est une macromolécule linéaire constituée de monomères de cellobiose, l'association de deux molécules de glucose liées par une liaison covalente en β1-4. Son degré de polymérisation, qui peut être extêmement variable, a un effet important sur ses propriétés et sa dégradabilité. La liaison en β1-4 permet des liaisons hydrogène intra et inter-chaînes (Figure 1), ce qui confère une grande résistance aux chaînes et surtout aux assemblages de chaînes, nommés microfibrilles (de 20 à 300 chaînes) et fibres (plus de 300 chaînes). Ces liaisons peuvent être à l'origine d'agencements ordonnés ou désordonnés, on parle alors de cellulose cristalline ou de cellulose amorphe. Le niveau de cristallinité de la cellulose a une influence sur l'accessibilité du substrat aux cellulases, et la cellulose amorphe présente des niveaux de dégradation 70% plus élevés que la fraction cristalline (Jeoh et al., 2007).

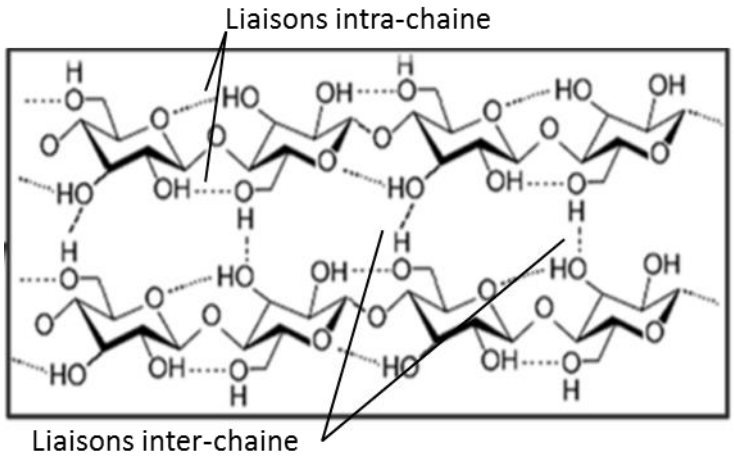


Figure 1 : Représentation schématique des liaisons hydrogène intra et inter-chaînes au sein d'une microfibrille de cellulose

La conversion des fibres de cellulose en monomères de glucose requière trois types de cellulases, ou glycoside hydrolases. Les endoglucanases permettent hydrolyser des liaisons osidiques à l'intérieur des chaînes de cellulose, les exoglucanases ou cellobiohydrolases libèrent du cellobiose (un dimère de glucose) à partir des extrémités réductrices libres, et enfin les β-glucosidases hydrolysent le cellobiose en deux molécules de glucose (Figure 2).

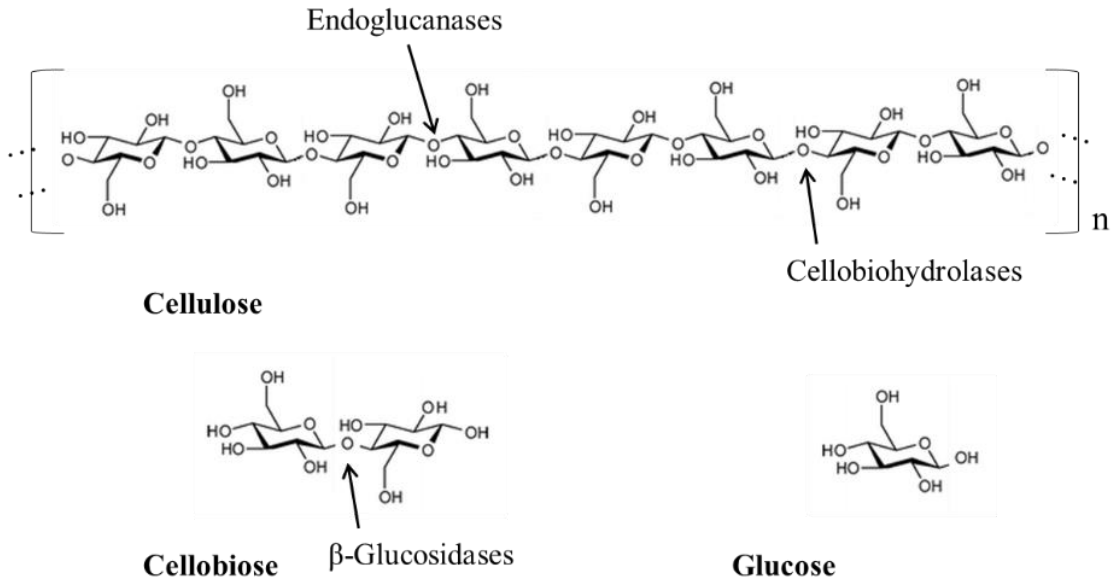


Figure 2 : Représentation schématique de la cellulose et des enzymes nécessaires à son hydrolyse.

I.1.2 Fraction hémi-cellulosique

L'hémicellulose diffère de la cellulose par son hétérogénéité biochimique. En effet, elle n'est pas constituée d'un monomère unique mais est un assemblage d'hétéropolymères constitués de pentoses (xylose, arabinose), d'hexoses (mannose, glucose, galactose) et d'acides glucuroniques. Les proportions de chaque composant peuvent varier d'une lignocellulose à l'autre. Alors que les bois tendres sont plutôt composés de glucomannanes, l'hémicellulose des plantes herbacées est souvent dominée par le xylose qui forme des chaînes liées en β1-4 appelées xylan. La composition plus hétérogène de l'hémicellulose et son organisation en chaînes latérales courtes, d'une plus grande accessibilité (Laureano-Perez et al., 2005), en font une molécule plus facilement hydrolysable que la cellulose. Néanmoins, du fait de son hétérogénéité, son hydrolyse requière une plus grande diversité d'enzymes, capables de s'attaquer aux différentes combinaisons de monomères et de liaisons.

La chaîne principale, composée de xylan dans le cas des plantes herbacées, peut être attaquée à l'intérieur de la chaîne par des endoxylanases, et les extrémités de chaînes peuvent libérer du xylobiose avec l'action d'exoxylanases. Les chaînes latérales doivent être également hydrolysées pour permettre l'accès au squelette de xylan. En fonction de la composition de ces chaînes latérales, différents types d'enzymes peuvent être impliquées, dont les plus fréquentes α-arabinofuranosidases, acetylxylan esterases ou α-glucuronidases. Une fois des dimères libérés, ceux-ci sont alors hydrolysés en monomères par des enzymes spécifiques, dont des β-xylosidases pour le xylobiose (Figure 3).

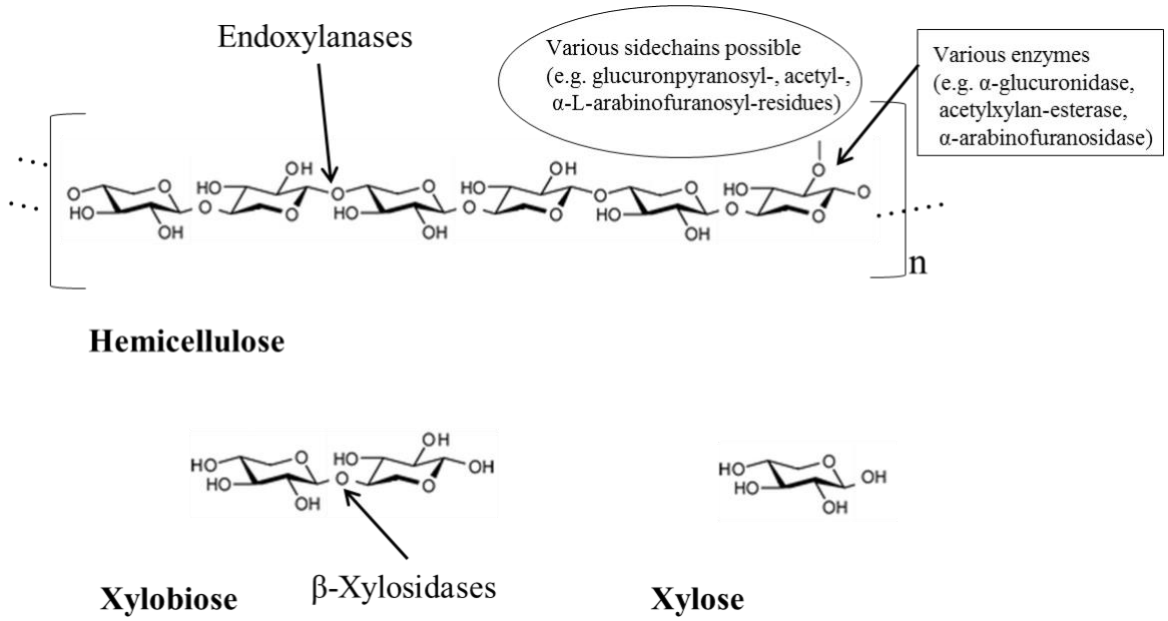


Figure 3 : Représentation schématique de l'hémicellulose et des enzymes nécessaires à son hydrolyse.

I.1.3 Lignine

La lignine est un polymère de molécules à cycles aromatiques dérivés de la phénylalanine, un acide aminé aromatique, comme les alcools p-coumaryliques, coniféryliques ou sinapyliques. Hormis cette composition à base de 3 principaux dérivés aromatiques, il n'existe pas de réelle unité dans ce qu'on regroupe sous le terme de lignine, les lignines pouvant être très variables entre espèces et même au sein d'une même espèce. Leurs fonctions principales sont de conférer imperméabilité, résistance mécanique et protection contre les attaques microbiennes. Sa structure complexe et variable, son hydrophobicité et sa très forte stabilité du fait de ses cycles aromatiques et la variété des liaisons qu'elle peut établir en font un obstacle très fort à l'hydrolyse des fibres de lignocellulose qui en sont imprégnées (Hendriks and Zeeman, 2009).

De fait, la lignine est difficilement dégradable pour la plupart des microorganismes, à l'exception notable de champignons (*Ascomycètes* et *Basidiomycètes* principalement). Ceux-ci sécrètent des enzymes, regroupées sous le nom de ligninases, classifiées principalement en phénol-oxydases (ou laccases), peroxydases à hème et enzymes accessoires.

Les laccases sont une famille d'oxydoréductases dont le mode d'action repose sur le couplage entre la réduction d'O_2 en H_20 avec l'oxydation de substrats comme les phénols ou la lignine. Les laccases connues partagent une structure tridimensionnelle commune en 3 domaines, le site actif contenant des ions cuivre, intervenants centraux de la réaction redox.

Les peroxydases à hème se divisent en plusieurs catégories, lignine peroxydases, manganèse-peroxydases et versatile peroxydases. Elles sont toutes H_2O_2 dépendantes et sont capables d'oxyder des substrats aromatiques phénoliques ou non phénoliques. L'optimum d'activité de la plupart d'entre elles se situe à un pH très acide.

Les enzymes accessoires ne sont pas directement impliquées dans la dégradation de la lignine mais interviennent dans des activités nécessaires en amont de l'action des ligninases, ou permettant de réduire les composés dérivés de la lignine. Elles incluent notamment des oxydases permettant la production de peroxyde d'hydrogène dont sont dépendantes les peroxydases.

I.1.4 Structure

La lignocellulose résulte donc du mélange, en proportions variables selon l'espèce de plante et les tissus végétaux, de ces trois composants majeurs mais également d'autres composants minoritaires (pectines, protéines). Leur organisation complexe est mise en place lors de la construction de la paroi végétale et peut varier selon les différents tissus végétaux (parenchymes, tissus vasculaires, tissus de soutien...). L'entrecroisement des fibres et la présence d'imprégnations de lignine liées de manière covalente aux autres composants (Figure 4) confèrent à la lignocellulose des propriétés de résistance supérieures à celles des composants pris individuellement, et en font un des polymères biologiques les plus résistants. Ces propriétés sont fondamentales pour la plante puisqu'elles permettent la turgescence cellulaire et donc le port dressé de la plante, mais sont également primordiales pour l'imperméabilité ou la défense face aux agressions extérieures.

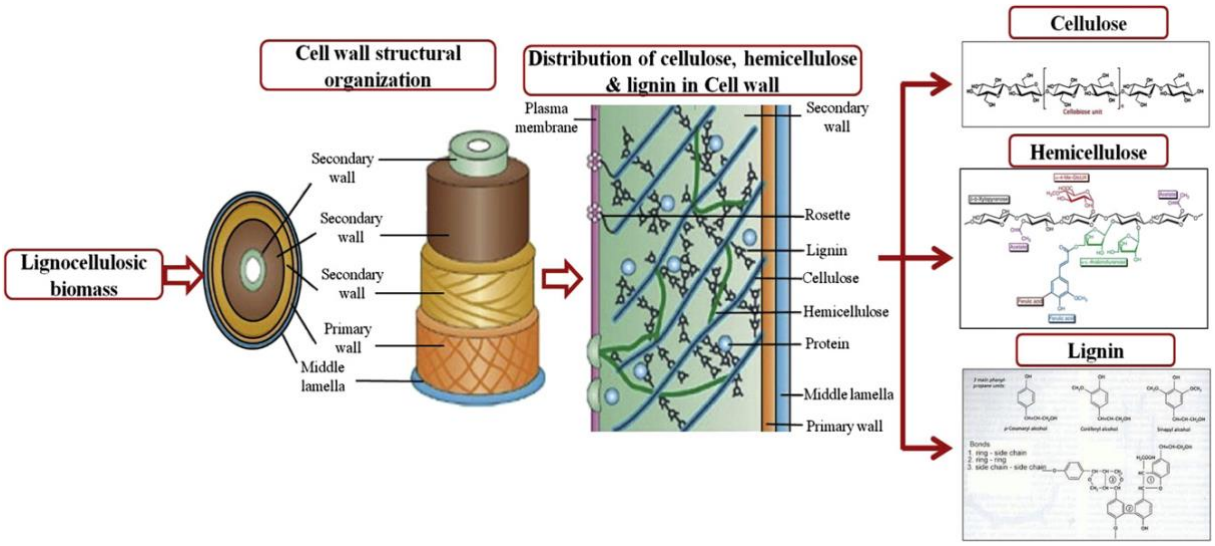


Figure 4 : Structure et organisation des composants de la lignocellulose (d'après Menon et al. 2012) Les propriétés de résistance des différentes molécules composant la lignocellulose (cellulose, hémicellulose et lignine), leur structure interne, les nombreuses liaisons covalentes ou faibles entre elles et la enfin leur organisation confère à la structure sa rigidité.

I.2. Procédés de valorisation actuels

Il existe différentes approches pour la valorisation des déchets végétaux. La plus ancienne, le compostage, a de loin précédé les connaissances en microbiologie, permet la production de compost, réutilisable en agriculture. Les approches plus récentes visent à produire de l'énergie (méthanisation) ou des molécules d'intérêt industriel (synthons) pour diverses applications telles que les plateformes des sucres ou des carboxylates.

I.2.1 Compostage et méthanisation

Le compostage est un processus biologique ressemblant beaucoup à la décomposition des sols, ou humification, qui existe en milieu naturel. Il consiste en la dégradation la partie carbonée des déchets végétaux pour en récupérer les éléments minéraux qui sont alors utilisés comme amendement pour reconstituer les sols agricoles. Il est fréquemment utilisé à l'échelle individuelle en conditions peu contrôlées, mais est aussi développé à grande échelle sur des plateformes de compostage industrialisées.

Le processus est réalisé en conditions aérobies et en présence d'eau. En présence d'oxygène, la dégradation des matières carbonées par respiration aboutit à la production de CO_2. L'activité microbienne associée à la respiration peut amener à des températures élevées, jusqu'à 70°C, ce qui explique l'existence et la présence de bactéries thermophiles capables de

dégrader la lignocellulose. Lors des différentes phases du compostage, on observe une alternance entre bactéries mésophiles qui débutent la dégradation, puis sont remplacées par des thermophiles lors de la montée en température. En fin de dégradation, la quantité de substrat facilement utilisable diminue et les microorganismes thermophiles disparaissent en faveur de bactéries mésophiles qui réalisent l'humification. Tout le processus étant réalisé en présence d'oxygène, il repose sur la conversion de matière organique en CO_2 et il est impossible de récupérer de l'énergie autrement que sous forme de chaleur. Il n'est pas nécessaire d'utiliser un inoculum exogène, le bon déroulement du compostage reposant principalement sur l'aération du substrat.

La méthanisation se produit naturellement dans certains environnements (sols non oxygénés, marais ou systèmes digestifs animaux). Elle consiste en une fermentation, c'est-à-dire la digestion anaérobie de la matière organique. En l'absence d'oxygène, la respiration est impossible et l'hydrolyse de la lignocellulose par des bactéries anaérobies aboutit à la production d'acides gras volatils, de dioxyde de carbone et d'hydrogène. Ceux-ci sont alors transformés en méthane par des Archées méthanogènes. Différentes applications existent à la méthanisation, dont les applications agricoles et le traitement des déchets ménagers, dynamisées par la recherche d'alternatives aux énergies fossiles.

La méthanisation permet de limiter la production de CO_2 et de transformer le substrat lignocellulosique en une molécule très simple mais énergétique, le méthane. En industrie, il est cependant intéressant de pouvoir récupérer des intermédiaires de la digestion anaérobie, que ce soit les sucres libérés par l'hydrolyse, ou les acides gras volatils produits par l'acétogénèse.

I.2.2 Plateforme des sucres

La plateforme des sucres est l'un des procédés de valorisation les plus largement étudiés. Elle consiste en la production de sucres en C5 ou C6 destinés dans un deuxième temps à la production d'éthanol ou de butanol principalement. Afin d'éviter les processus d'acétogénèse et de méthanisation qui vont de pair avec une hydrolyse microbienne, dans cette plateforme la lignocellulose est hydrolysée par voie enzymatique uniquement, en l'absence de micro-organismes. Les sucres monomériques libérés sont ensuite fermentés en

utilisant des souches sélectionnées, souvent modifiées génétiquement, du type de la levure *Saccharomyces cerevisiae*.

L'hydrolyse est effectuée à l'aide de cocktails enzymatiques commerciaux qui contiennent en général les enzymes excrétées d'un ou plusieurs micro-organismes, la plupart du temps des champignons. De nombreuses études se sont intéressées à la production de cocktails d'enzymes cellulolytiques, puis hémicellulolytiques. Le micro-organisme le plus étudié est *Tricoderma reesei*, un champignon capable de produire de hauts niveaux d'enzymes cellulolytiques. D'autres organismes ont également été étudiés pour leur production de cellulases et hémicellulases, comme *Aspergillus, Cellulomonas, Clostridium* ou *Trametes.* Certains travaux s'attaquent au problème d'une dégradation au moins partielle de la lignine et la production d'enzymes ligninolytiques de champignons (*Phanerochaete chrysosporium)* ou de bactéries (*Streptomyces*) ont également été étudiées (Menon and Rao, 2012). Des efforts conséquents ont été déployés pour améliorer les cocktails enzymatiques en découvrant de nouvelles enzymes, y compris par des approches méta-omiques ou en augmentant leurs performances par ingénierie enzymatique. Cependant, l'hydrolyse reste l'étape limitante à la fois en termes de rendements, mais aussi de coûts. 10 à 25 grammes de protéines sont nécessaires par kilogramme de biomasse à traiter, et le coût de production des enzymes représente plus de la moitié du prix de l'éthanol final (Klein-Marcuschamer et al., 2012). La digestion du substrat n'est que partielle, les cocktails enzymatiques même les plus complets ne présentant pas une diversité enzymatique aussi grande que celle des communautés microbiennes qui réalisent l'hydrolyse en environnement naturel. Les méthodes de prétraitement physiques ou chimiques permettent d'augmenter la digestibilité du substrat utilisé, mais celle-ci reste tout de même limitée.

Enfin, il est nécessaire pendant tout le processus d'éviter la présence de micro-organismes qui consommeraient les sucres produits lors de l'étape d'hydrolyse, ou entreraient en compétition avec la souche utilisée pendant la fermentation. Ce besoin de conditions stériles a un impact important sur le coût du procédé.

I.2.3 Plateforme des carboxylates

Afin d'éviter le recours à des cocktails enzymatiques et des conditions stériles, il est possible, de la même façon qu'en méthanisation, d'utiliser directement des communautés

microbiennes lignocellulolytiques pour la digestion de la biomasse. En contrepartie d'une consommation limitée de substrat pour la croissance microbienne. S'il est impossible d'envisager obtenir des communautés microbiennes réalisant uniquement l'étape d'hydrolyse, il est possible d'empêcher la dernière phase du procédé, la méthanogénèse par les Archées, leurs conditions de culture étant relativement spécifiques, notamment en termes d'acidité. Cette digestion anaérobie sans production de méthane est appelée plateforme des carboxylates (Agler et al., 2011), ceux-ci constituant le produit final. Elle a en commun avec la méthanisation (ou plateforme biogaz) la plupart de ses étapes, hydrolyse, acidogénèse et acétogénèse. Les acides gras à courtes chaînes (carboxylates) en sont le produit d'intérêt, et peuvent une fois purifiés servir soit directement en chimie industrielle (acétate, butyrate) ou être convertis en bioplastiques comme les polyhydroxyalcanoates (Torella et al., 2013). Les rendements observés sont les plus élevés comparés aux autres plateformes, même si une bonne marge de progression existe jusqu'aux rendements théoriques limites (Holtzapple and Granda, 2009).

La plateforme des carboxylates est directement couplée à la production d'alcools via l'injection d'hydrogène au sein du procédé MixAlco. L'étude économique de celui-ci montre qu'au-delà des enjeux écologiques, sa rentabilité financière est possible puisque le prix du carburant produit est comparable à celui d'un carburant fossile (Granda et al., 2009). Associée à ses avantages par rapport aux autres plateformes (pas d'asepsie, structures peu coûteuses utilisables, adaptable à une grande gamme de substrats, pas d'enzymes à ajouter, pas de micro-organismes génétiquement modifiés, pas d'incidents de contamination…), ces bons rendements font de la plateforme carboxylate le meilleur procédé actuel de conversion de la biomasse lignocellulosique.

Pour optimiser le procédé, de nombreux leviers existent, dont les paramètres de culture, le prétraitement du substrat mais également l'apport d'inoculum exogène. L'inoculation par une communauté enrichie en micro-organismes lignocellulolytiques offre plusieurs avantages. Elle permet d'augmenter la biomasse lignocellulolytique dès le début de la digestion et donc de l'accélérer en s'affranchissant en partie de la croissance microbienne. Elle permet aussi l'apport de communautés plus efficaces et plus rapides pour la dégradation que les communautés endogènes. Plusieurs équipes travaillent à la constitution de communautés bactériennes aux capacités hydrolytiques et à la stabilité élevée, qui seront détaillées en partie II.3.2.

I.3. Prétraitements

Sous sa forme native, la lignocellulose est récalcitrante à l'hydrolyse enzymatique et microbienne. De ce fait, elle est lentement biodégradée et son niveau de conversion est faible, ne dépassant pas les 40% (Barakat et al., 2013). Plusieurs propriétés inhérentes au matériel lignocellulosique limitent son hydrolyse. En particulier, la lignine, du fait de son imperméabilité et des multiples liaisons qui l'associent à l'holocellulose, est considérée comme le 1er facteur limitant. En plus d'être une barrière physique résistante aux agressions chimiques et biologiques, les effets négatifs de la lignine sur la bioconversion de la lignocellulose incluent également l'adsorption non-spécifique des enzymes hydrolytiques, des interférences avec les enzymes cellulolytiques qui s'associent de manière non productive aux complexes lignine-carbohydrates, et la libération de dérivés de la lignine toxiques pour les microorganismes (Agbor et al., 2011). Mais la récalcitrance de la biomasse n'est pas uniquement due à sa teneur en lignine, elle est également le fait de propriétés structurelles, telles que le degré de polymérisation, la solidité des fibres (nombre de liaison et organisation) et la cristallinité de la cellulose, qui limitent l'étendue et la vitesse d'hydrolyse de la lignocellulose. Enfin, les critères de porosité (ou surface spécifique) et d'accessibilité de la cellulose aux enzymes hydrolytiques jouent un rôle majeur dans la biodégradation de la lignocellulose par les microorganismes et leurs enzymes (Hendriks and Zeeman, 2009).

I.3.1 Efficacité des prétraitements de la biomasse

L'application de prétraitements sur la biomasse lignocellulosique en amont de son hydrolyse enzymatique ou microbienne est une étape essentielle permettant de modifier les liaisons et la structure supramoléculaire de la matrice cellulose-hémicellulose-lignine. Cela permet d'augmenter l'accessibilité et la biodégradabilité de l'holocellulose (Jeoh et al., 2007). Ces dix dernières années, de nombreux articles scientifiques ont été publiés sur les applications et les effets de prétraitements de la lignocellulose. Les objectifs du prétraitement sont la destruction de la barrière ligneuse, l'augmentation de la porosité et de l'accessibilité au substrat, et la diminution de la cristallinité et du degré de polymérisation. La Figure 5 est une représentation schématique des effets des prétraitements. De manière globale, pour être efficaces, les prétraitements doivent satisfaire à quatre critères : (i) améliorer l'hydrolyse de la lignocellulose en sucres simples, (ii) limiter la dégradation des carbohydrates, (iii) éviter la

formation de sous-produits inhibiteurs à l'action des enzymes ou à la fermentation microbienne, (iv) être économiquement viables (Chandra et al., 2012).

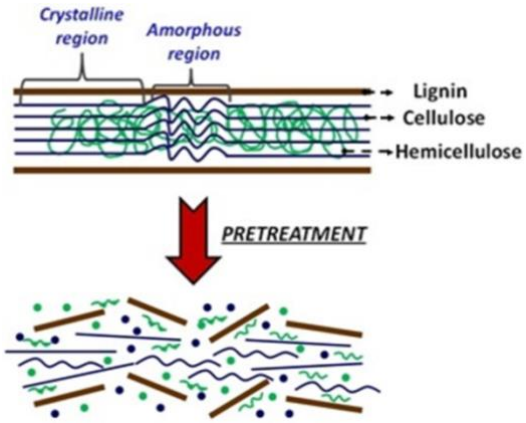


Figure 5 : Représentation schématique des effets du prétraitement sur la lignocellulose. ©Mora-Pale

I.3.2 Méthodes de prétraitement

De nombreuses méthodes de prétraitements ont été étudiées et de nombreuses autres sont encore en développement. La comparaison de ces différentes méthodes n'est pas chose aisée car la qualité d'un prétraitement dépend de l'optimisation des facteurs coût et efficacité. Or l'évaluation de la rentabilité économique est complexe car elle implique des coûts de procédés amont et aval, des investissements en capitaux, ainsi que des systèmes de recyclage des déchets et produits chimiques qui sont variables (Agbor et al., 2011). D'autre part, l'efficacité des différentes technologies varie également en fonction du substrat lignocellulosique traité et peut-être évaluée différemment en fonction du produit final recherché (sucres simples, acides gras volatils, éthanol, méthane, etc.).

I.3.3 Méthodes physiques de prétraitement

Les prétraitements mécaniques de coupe, mise en copeaux ou broyage de la biomasse lignocellulosique est une étape indispensable à son hydrolyse enzymatique ou microbienne. Ils permettent une réduction de la taille particulaire allant de 50 mm jusqu'à des diamètres inférieurs à 60 µm. Leurs principaux effets sont une augmentation de la surface spécifique d'échange (et donc une meilleure accessibilité au substrat) et une diminution du degré de polymérisation. Pour ce qui est de la cristallinité, différentes études donnent des résultats contradictoires (Agbor et al., 2011; Barakat et al., 2014). Globalement, la réduction mécanique des particules améliore l'hydrolyse du substrat : diminution du temps (-23 à -59%)

et augmentation du rendement d'hydrolyse (+5 à +25%) ; les rendements de fermentation sont également améliorés pour la production de biogaz, bioéthanol et biohydrogène (Hendriks and Zeeman, 2009). Par contre, ce sont des procédés à forte demande énergétique et donc coûteux.

Il existe également d'autres prétraitements physiques moins courants qui peuvent être appliqués à la lignocellulose. Les procédés thermiques permettent de solubiliser l'hémicellulose puis la lignine en exposant le substrat à des températures supérieure à 150-180°C. Ces méthodes présentent cependant un risque élevé de condensation/précipitation de la biomasse par la lignine solubilisée (diminution de l'accessibilité du substrat) et de production de produits de dégradations toxiques ou inhibiteurs : composés phénoliques et hétérocycliques (Furfural et 5-hydroxy-methyl furfural ou 5-HMF) (Hendriks and Zeeman, 2009). Les procédés d'irradiation aux rayons gamma permettent une augmentation de la surface accessible et une diminution de la cristallinité de la cellulose grâce à la destruction des liaisons β-1,4. La technique est cependant trop coûteuse à l'échelle industrielle et présente des risques pour la santé humaine et l'environnement.

I.3.4 Méthodes chimiques de prétraitement

Parmi les prétraitements de type chimique, les traitements de type alcalin font l'objet de nombreuses publications. Le procédé consiste à traiter la biomasse lignocellulosique par une solution de soude (NaOH), de chaux ($Ca(OH)_2$) ou d'ammoniac aqueux ($NH_3.H_2O$) à une concentration donnée, pendant un temps plus ou moins long. L'action des composés basiques a pour effet de couper les liaisons lignine-carbohydrates et d'éliminer les groupements acétyles et acides uroniques de l'hémicellulose, permettant ainsi d'améliorer l'accessibilité au substrat. Les études rapportent également un gonflement de la lignocellulose traitée avec une solution alcaline et donc une augmentation de la surface spécifique interne du substrat. Ce traitement induit ainsi une diminution du degré de polymérisation et de la cristallinité, et une solubilisation partielle de l'hémicellulose et de la lignine. Les réactifs sont peu coûteux et leurs effets sur la lignocellulose permettent d'améliorer efficacement les étapes d'hydrolyse et de fermentation. Par contre, le traitement est surtout efficace sur les substrats à faible teneur en lignine (résidus agricoles) ; et selon la sévérité du procédé, il a des risques de pertes de carbone sous forme du CO_2 et de formation de produits de dégradation toxiques ou inhibiteurs (Furfural et 5-HMF) (Agbor et al., 2011; Chandra et al., 2012; Hendriks and Zeeman, 2009).

Les traitements acides permettent également d'améliorer la digestibilité de la lignocellulose en modifiant les propriétés physicochimique du substrat. L'acide sulfurique (H_2SO_4), l'acide chlorhydrique (HCl) ou l'acide acétique (CH_3COOH) sont utilisés en solution diluée pour hydrolyser l'hémicellulose (en particulier le xylan) en monomères de sucres et augmenter de ce fait l'accessibilité du substrat cellulosique. En plus de la libération de composés toxiques ou inhibiteurs, et la nécessité éventuelle de neutraliser la biomasse avec une base (NaOH) pour sa fermentation, les prétraitements acides ont l'inconvénient de la gestion de solutions corrosives et la mise en place d'un recyclage des produits chimiques pour un procédé économiquement viable (Agbor et al., 2011).

I.3.5 Méthodes physico-chimiques

La technique d'explosion à la vapeur (Steam Explosion) consiste à exposer la lignocellulose à de fortes pressions (0.6-0.8 MPa) et températures (160-240°C) dans une atmosphère saturée en vapeur d'eau pendant quelques minutes. Ensuite, une dépressurisation rapide permet de déstructurer la matrice lignine-carbohydrates et solubiliser l'hémicellulose dans la phase liquide. Le prétraitement permet d'augmenter l'accessibilité et la réactivité de la cellulose aux enzymes hydrolytiques avec un procédé à faible demande énergétique et sans addition de produits chimiques. La méthode entraîne par contre une perte non négligeable de la fraction xylan et donc de pentoses potentiellement valorisables ainsi que la libération de composés toxiques ou inhibiteurs. Le risque de condensation de la biomasse et de diminution de l'accessibilité du substrat est cependant élevé (Sun and Cheng, 2002).

Les traitements à l'eau chaude liquide LHW (Liquid Hot Water) ont des effets similaires aux traitements à la vapeur avec l'hydrolyse de l'hémicellulose et la délignification du substrat. Par contre, les faibles températures et pressions limitent la libération de composés de dégradation toxiques ou inhibiteurs, et les grandes quantités d'eau permettent de solubiliser une plus grande fraction de sucres qui sont en contrepartie fortement dilués.

Les systèmes AFEX (Ammonia Fiber EXplosion) ou ARP (Ammonia Recycle Percolation) combinent l'action du catalyseur alcalin NH_3 avec celle de procédés physiques à haute pression ou percolation. Les répercussions sur le substrat sont semblables à ceux obtenus par traitement alcalin avec de surcroit, une altération de la lignine qui facilite sa solubilisation tout en limitant la production d'inhibiteurs ou de produits toxiques. Certaines études rapportent une forte augmentation de la réactivité de la cellulose lors de l'hydrolyse

enzymatique de la biomasse prétraitée par AFEX ou ARP avec des rendements proches des rendements théoriques et une forte augmentation de la production d'éthanol en fermentation (Alizadeh et al., 2005; Kim et al., 2006).

Les prétraitements oxydatifs WOP (Wet Oxydative Pretreatment) utilisent le peroxyde d'hydrogène (H_2O_2) ou l'acide peroxyacétique (CH_3CO_3H) pour délignifier efficacement la biomasse à forte teneur en lignine. La méthode a l'inconvénient d'être peut sélective et entraine la perte d'une fraction importante des sucres de l'holocellulose. De plus, le risque de la formation d'inhibiteurs par oxydation des composés aromatiques est élevé (Hendriks and Zeeman, 2009).

La méthode Organosolv consiste à employer un mélange de solvant organique et solvant aqueux pour solubiliser l'hémicellulose et extraire la lignine. Le prétraitement est très sélectif et permet de séparer la lignocellulose en trois fractions, avec la fraction cellulose quasiment pure, l'hémicellulose en phase aqueuse et la lignine en phase solide. L'inconvénient majeur de cette technique est le coût du procédé : les solvants organiques, en plus d'être chers (système de recyclage nécessaire pour limiter les pertes), présentent un risque pour la santé humaine et l'environnement, et doivent être utilisés dans des installations confinées (Agbor et al., 2011).

Les solvants ioniques IL (Ionic Liquid) sont des sels en solution qui entraînent la déconstruction de la structure tridimensionnelle de la lignocellulose par l'entrée en compétition des ions avec les liaisons hydrogène. Ce type de prétraitement favorise le fractionnement de la biomasse avec la possibilité de récupérer la fraction cellulose par l'ajout d'eau, d'éthanol ou d'acétone. L'approche IL pour le prétraitement de la biomasse est relativement nouvelle et fait encore l'objet d'investigations (Agbor et al., 2011).

I.3.6 Autres méthodes

Dans le cadre de la valorisation de ressource lignocellulosique comme source de carbone, les aspects écologiques des prétraitements ne peuvent pas être négligés, notamment leur coût en énergie et en eau. De nouveaux prétraitements « eco-friendly » ont été étudiés récemment, reposant sur un traitement chimique par imprégnation du substrat, limitant ainsi le volume d'eau utilisé (Barakat et al., 2014) combiné au prétraitement mécanique par broyage

ou extrusion. Cette technique a été testée avec un traitement chimique à la soude, et peut être associée ou non à un prétraitement mécanique simultané.

Il existe également des méthodes de prétraitements biologiques qui font appel aux capacités de certains champignons à dépolymériser la lignocellulose. Ces organismes appartenant au règne des *Fungi* ou *Mycota* produisent des enzymes (lignine peroxydases, laccases, polyphénols oxydases) capables de dégrader la lignine, l'hémicellulose et les polyphénols (Canam et al., 2013). La délignification de la biomasse est efficace et sélective permettant d'augmenter l'accessibilité du substrat et les vitesses de production de biohydrogène en fermentation (Magnusson et al., 2008). L'inconvénient majeur des prétraitements biologiques est la lenteur du procédé (10 jours à plusieurs semaines) avec des conditions contrôlées et une perte du carbone consommé par les champignons (Agbor et al., 2011).

Enfin, les méthodes d'extrusion font intervenir un broyage par une extrudeuse (vis sans fin). Elles sont peu coûteuses en énergie comparé à d'autres méthodes mécaniques, et on l'avantage de permettre un contrôle de la température à des valeurs basses, ce qui évite la dégradation des carbohydrates et l'oxydation de la lignine, souvent libératrice d'inhibiteurs de fermentation (Lin, 2013). Elle permet également d'appliquer par injection en cours d'extrusion des prétraitements chimiques ou enzymatiques (extrusion réactive) et présente de nombreux avantages qui en font une méthode à l'avenir prometteur : elle est pratique et facile à utiliser à une échelle industrielle, permet de mettre en place un prétraitement continu, n'induit pas de perte de matière (Lin, 2013).

II. Dégradation naturelle de la lignocellulose

Les végétaux étant quasiment les seuls autotrophes au carbone dans les écosystèmes terrestres, tous les organismes hétérotrophes dépendent, à un moment ou un autre de leur chaîne alimentaire, de la digestion de lignocellulose. Dégradation, méthanisation, humification ou minéralisation de la biomasse lignocellulosique sont des processus naturels réalisés par des micro-organismes qui font partie du cycle du carbone ; ils ont lieu notamment dans les sols. Les animaux herbivores, se nourrissant majoritairement de plantes, sont étonnement dépourvus du matériel enzymatique nécessaire à l'hydrolyse de la lignocellulose. Si certains d'entre eux ont une production de un nombre restreint d'enzymes lignocellulolytiques, la digestion complète de la lignocellulose repose principalement sur l'intervention de flores digestives microbiennes. Ainsi, la transformation de la biomasse lignocelulosique a lieu principalement grâce au métabolisme des microorganismes présents dans les sols ou les systèmes digestifs. L'étude du fonctionnement de ces écosystèmes, capables d'atteindre de niveau de dégradation inégalés en industrie permet une meilleure compréhension des mécanismes de transformation de la lignocellulose, mais également de s'en inspirer ou d'utiliser leur potentiel en biotechnologie. Ici, nous porterons notre attention à deux systèmes digestifs, celui des mammifères d'élevage et celui d'insectes, particulièrement des termites, capables de dégrader des structures végétales très lignifiées comme le bois.

II.1. Digestion de la biomasse chez les mammifères ruminants

Les mammifères sont le groupe animal le plus étudié du fait de leur importance économique en élevage et alimentation humaine. Il existe cependant, au sein même des mammifères herbivores, une très grande variété de systèmes digestifs, mais le système digestif le mieux connu chez les mammifères est celui des ruminants. Leur particularité est de posséder un estomac à quatre compartiments. Trois d'entre eux correspondent à un « pré-estomac », le rumen, le réseau et le feuillet, précédant la caillette, l'équivalent de l'estomac chez les Mammifères monogastriques. L'essentiel de la digestion a lieu dans le rumen, le premier compartiment mais également le plus volumineux (jusqu'à 200L chez un bovin adulte). Elle est essentiellement microbienne, même si la salive sécréter apporte également quelques enzymes, mais le processus de rumination (régurgitation suivie de mastication) joue un rôle non négligeable de mélange et de fractionnement des aliments, ce qui facilite l'action des micro-organismes et leurs enzymes.

L'hydrolyse de la lignocellulose en sucres simples est réalisée par les enzymes microbiennes. Le rumen étant un environnement anaérobie, ces sucres sont alors fermentés en acide pyruvique puis acides gras volatils, principalement acétate, propionate et butyrate. Ceux-ci sont alors absorbés au niveau des papilles épithéliales de la paroi ruminale (~90%) ou du feuillet et métabolisés soit dans la paroi du rumen soit par le foie. Le butyrate est métabolisé en β-hydroxy butyrate, le propionate est transformé en lactate ou en oxaloacétate. L'acétate passe intégralement dans le sang et est métabolisé dans le foie en entrant dans le cycle de Krebs sous forme d'acétyl-CoA.

La fermentation dans le rumen s'accompagne d'une production de méthane (jusqu'à 8% du total des calories), qui représente une perte d'énergie pour l'animal puisqu'il est évacué par éructation. Cette production peut provenir de deux voies, par consommation de CO_2 et H_2 (voie principale) ou par consommation de carboxylates. Le CO_2 provenant principalement de la décarboxylation de l'acide pyruvique en acétate, les régimes favorisant une production de propionate et diminuant celle d'acétate permettent de limiter la production de méthane.

Le rumen est un fermenteur microbien dont la température, la concentration et l'acidité sont régulés (40°C, ~350mosm.L^{-1}, pH optimal autour de 6). C'est un écosystème microbien très riche composé de bactéries (10^{10} à 10^{11}/g), d'archées (10^8 à 10^{10}/g) et d'eucaryotes (10^2 à 10^4/g champignons et 10^4 à 10^6/g protozoaires). Les protozoaires sont principalement des espèces ciliées anaérobies, dominées par le genre *Entodinium*. Ils sont assez sensibles à la diminution de pH en cas d'acidose mais contribuent à l'éviter en métabolisant l'acide lactique. Ils ne sont pas tous cellulolytiques mais participent à la digestion par ingestion de particules par phagocytose et ont également un rôle de régulation des populations bactériennes. Les champignons anaérobies présentent une biologie et un métabolisme inhabituels et produisent des cellulases, xylanases et esterases phénoliques. Ils sont principalement impliqués dans l'hydrolyse des fibres de la lignocellulose et sont capables de dégrader ou du moins d'altérer la lignine (Caporaso et al., 2010). Les archées, représentées principalement par les genres *Methanobacterium* et *Methanobrevibacter* (Janssen and Kirs, 2008) réalisent la méthanogénèse à partir de CO_2 et d'H_2 et ne présentent pas d'activité hydrolytique. Enfin, la fraction bactérienne sera discutée en partie II.3.1.

II.2. Insectes

Les Insectes sont la classe la plus diversifiée des organismes animaux, et une très grande partie d'entre eux sont humivores ou xylophages. Ceux-ci sont donc particulièrement intéressants pour leur capacité à digérer la lignocelulose même très lignifiée. En effet, de très nombreux insectes ont des régimes alimentaires constitués strictement de lignocellulose, que ce soit à l'état adulte, mais aussi très souvent larvaire. Des gènes codant pour des cellulases (notamment des familles de glycoside hydrolases GH5, GH9 et GH45) ont récemment été identifiés chez les termites, les cafards ou encore les phasmes (Cragg et al., 2015), mais les productions d'enzymes endogènes sont systématiquement complémentées par des flores microbiennes lignocellulolytiques.

II.2.1 Les Termites

Parmi les Insectes, les termites (Isoptères) sont parmi les plus abondants sur Terre. Ils sont présents sur tous les continents excepté l'Antarctique et peuvent représenter jusqu'à 90% de la biomasse d'insectes dans les régions tropicales et sous-tropicales (Bignell and Eggleton, 2000) et leur capacité à dégrader la lignocellulose leur donne une place centrale dans le cycle du carbone de ces régions.

Les termites ont pour points communs d'avoir une alimentation basée sur la lignocellulose, digérée en s'appuyant sur des micro-organismes symbiotiques, et de vivre en communautés eusociales. La plupart des espèces de termites sont organisées en trois castes : reproducteurs, soldats et ouvriers. Les reproducteurs (reines et rois) sont les plus rares et sont les seuls à produire des œufs. Ils sont ailés et peuvent fonder de nouvelles colonies à distance de leur nid d'origine, avant de perdre leurs ailes. Les soldats servent principalement à défendre le nid contre des intrus. Deux morphologies principales existent selon les genres. Les *Nasutitermes* par exemple présentent des soldats à trompe, capables de projeter une substance collante et corrosive, tandis que les genres *Termes* ont des soldats dont les mandibules sont beaucoup plus développées que celles des ouvriers. Enfin, les ouvriers ont des tâches variées comme la collecte de nourriture, l'alimentation des autres castes ou la construction du nid.

La classification phylogénétique définitive des termites est encore sujette à débat, mais certains regroupements (monophylétiques ou non) peuvent être faits sur la base de critères morphologiques, écologiques ou physiologiques. On peut distinguer quatre groupes, basés sur des critères morphologiques et d'habitudes alimentaires (Donovan et al., 2001). Même si la capacité de dégradation de la lignocellulose est une caractéristique commune à tous les termites (xylophages ou non), elle ne s'appuie toujours pas sur les mêmes mécanismes et les mêmes symbiotes. Les *Macrotermitinae,* appelés aussi termites champignonnistes, utilisent les capacités lignolytiques de champignons (*Termitomyces*) dans une symbiose externe, ou ectosymbiose. La termitière comporte des champignonnières où les ouvriers déposent le substrat grossièrement dégradé par mastication en meules aérées, et l'inoculent avec le champignon. Celui-ci pré-dégrade la lignocellulose, qui est ensuite consommée par les termites (Aanen et al., 2007). Ce cas particulier, faisant intervenir un champignon aérobie, est assez éloigné d'éventuelles applications en fermentation et ne sera pas davantage détaillé ici. Champignonnistes exclus, on distingue deux grand types de symbioses digestives, celle des termites inférieurs, la plus étudiée, et celle des termites supérieurs, actuellement moins connue.

II.2.2 Termites inférieurs

Les termites inférieurs présentent une flore digestive constituée de deux grandes composantes, l'une bactérienne et l'autre protiste (Flagellés), absente chez les termites supérieurs. Ces protistes, colonisant l'intestin postérieur (Figure 6), et sont responsables de l'hydrolyse complète de la cellulose, mais aussi de la libération de produits de fermentation qui sont absorbés par l'hôte (Breznak and Brune, 1994). L'hydrolyse de la biomasse est intracellulaire, les particules de bois étant internalisées par phagocytose. Sous un régime de bois, cette symbiose avec des Flagellés anaérobies est vitale pour l'hôte. En effet, l'élimination des flagellés (e.g. par oxygénation de l'intestin) conduit à la mort de faim du termite même si celui-ci continue à s'alimenter, sauf s'il est ré-inoculé avec des flagellés (Ebert and Brune, 1997) : la symbiose est donc essentielle pour la digestion de la lignocellulose. Les assemblages de flagellés peuvent être complexes (jusqu'à 19 espèces chez *Hodotermopsis japonica*) mais aussi très simples (trois espèces différentes seulement chez *Coptotermes formosanus*) (Bignell and Eggleton, 2000). Chaque espèce, morphologiquement

discernables, joue un rôle spécifique dans la digestion (Brune, 2014). Certaines espèces sont cellulolytiques, xylanolytiques et d'autres, de très petite taille, ne semblent pas phagocyter de particules mais participent à la digestion des substrats solubles. Le répertoire enzymatique de ces flagellés a changé pendant leur co-évolution : si les glycoside hydrolases de la famille 7 (GH7) semblent avoir été présentes chez les flagellés avant leur association avec les termites, d'autres familles comme les GH5, GH10 ou GH11 auraient été acquises par transfert horizontal depuis d'autres micro-organismes de la flore intestinale (Todaka et al., 2010).

Les bactéries sont également représentées dans la flore des termites inférieurs. Notamment, le phylum des *Spirochaetes* est souvent très abondant chez les termites xylophages, où ils peuvent compter pour la moitié des procaryotes. Contrairement au *Spirochaetes* qui sont souvent libres, la plupart des autres bactéries et archées sont soit associées à la paroi de l'intestin ou intégrées à des structures qui y sont rattachées, soit associées aux flagellés, à leur surface ou dans leur cytoplasme (Hongoh, 2011). Cette internalisation de bactéries constitue ainsi une symbiose complexe à trois types de partenaires.

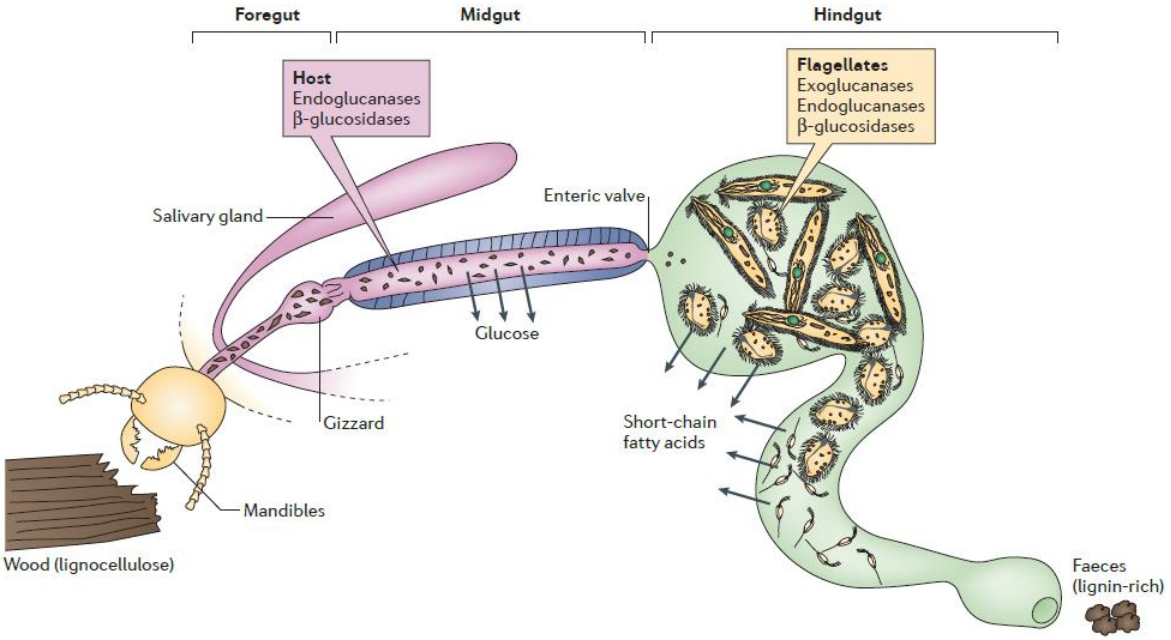


Figure 6 : schéma récapitulant la digestion chez les termites inférieurs (d'après (Brune, 2014)) **La digestion de la lignocellulose implique l'activité de l'hôte et de sa flore intestinale. Chez le termite inférieur, de petites particules de bois sont produites par les mandibules de l'hôte, additionnées d'enzymes produites par l'hôte et réduites par l'action mécanique du gésier. Les particules de bois, partiellement digérées, passent ensuite la valve entérique et pénètrent dans l'intestin postérieur où elles sont phagocytées et digérées intracellulairement dans des vacuoles digestives. Les produits de fermentation libérés sont absorbés par l'hôte au niveau de la paroi intestinale.**

II.2.3 Termites supérieurs

Les termites supérieurs (famille des Termitidae) ne présentent aucune flore de flagellés. La fonction de digestion de la lignocellulose est néanmoins maintenue par d'autres partenaires symbiotiques qui ne sont pas moins efficaces: les termites supérieurs comptent pour plus de 80% des espèces de termites et présentent des régimes alimentaires plus diversifiés que les termites inférieurs. La perte des flagellés s'accompagne ainsi de modifications anatomiques telles que l'élongation de l'intestin, compartimentation plus forte ou alcalinité augmentée de l'intestin antérieur. Elle s'accompagne aussi de modifications symbiotiques liées au régime alimentaire : la flore des termites xylophages est dominée par les *Spirochaetes* et *Fibrobacteres* tandis que celle des termites humivores est dominée par les Firmicutes (Bignell et al., 2010). Les communautés bactériennes semblent façonnées à la fois par la phylogénie de l'hôte et par son régime alimentaire (Mikaelyan et al., 2015a). Elle est également très dépendante du stade de développement des individus, et change considérablement entre stades larvaires et ouvriers adultes (Diouf et al., 2015).

La distribution des fonctions entre hôte et microbiote a longtemps été discutée. L'endothélium de l'intestin intermédiaire des termites supérieurs sécrète des cellulases (endoglucanases et glucosidases) qui permettent la digestion partielle de la cellulose (Tokuda et al., 2004). Par ailleurs, l'activité cellulase détectée dans l'intestin postérieur est très faible, et n'explique à elle seule les taux élevés de dégradation observés. Des récents travaux ont montré chez *Nasutitermes* qu'une activité cellulase très forte dans l'intestin supérieur était associée à la fraction particulaire ; telle activité n'était pas détectée dans les protocoles classiques de mesure dans le surnageant (Tokuda and Watanabe, 2007). L'hydrolyse de la cellulose chez les termites supérieurs comme inférieurs est donc partagée entre l'hôte (hydrolyse partielle de la cellulose amorphe) et le microbiote microbien (hydrolyse de la cellulose cristalline et/ou protégée par les autres composants de la lignocellulose).

Dans le métagénome de l'intestin postérieur des *Nasutitermes,* les gènes assignés *Fibrobacteres* sont extrêmement représentés, codant pour des protéines très majoritairement porteuses de domaines de liaison à la cellulose (Warnecke et al., 2007). Les membres du genre *Fibrobacter* sont très connus dans la flore ruminale, mais ils manquent de cellulases solubles ou d'autres protéines caractéristiques des cellulosomes des *Clostridia*. Ils pourraient

cependant avoir une activité conséquente passant par une forte interaction avec le substrat et par la colonisation des particules de bois (Mikaelyan et al., 2014).

II.3. Études de la diversité associée à la dégradation de la lignocellulose

La lignocellulose est présente dans de nombreux écosystèmes où elle est dégradée et sert de source de carbone : sédiments marins, sols, compost ou encore systèmes digestifs animaux. Les communautés microbiennes impliquées ne sont pas uniquement bactériennes. Les champignons y jouent parfois un rôle majeur, notamment dans les écosystèmes forestiers (Dashtban et al., 2009). Cependant, dans un contexte de valorisation de biomasse lignocellulosique, les communautés aérobies ne sont pas un bon modèle puisqu'en présence d'oxygène la dégradation est principalement orientée vers la production de biomasse microbienne et de CO_2, alors que les produits intéressants en industrie sont les produits de fermentation (acides gras volatils notamment). Seuls quelques champignons sont connus pour croître en conditions anaérobies et produire des enzymes lignocellulolytiques (cellulases, xylanases et phénol-hydrolases), mais ils n'ont été détectés que dans un nombre limité d'écosystèmes (intestins de ruminants principalement (Gordon and Phillips, 1998; Cheng et al., 2009)). Dans la majeure partie des autres cas, la dégradation de la lignocellulose sous conditions anaérobies est assurée en grande partie par la composante bactérienne des communautés microbiennes.

II.3.1 Exemple en système naturel : le rumen

Le rumen est un des écosystèmes naturels dégradant la lignocellulose qui est le plus étudié, à la fois pour des raisons agronomiques et biotechnologiques. En effet, la compréhension du fonctionnement de sa microbiologie peut permettre d'améliorer la digestion et l'efficacité alimentaire du bétail, ce qui est intéressant agronomiquement (Jouany, 2006), mais peut aussi offrir de nouvelles possibilités en biotechnologies avec la découverte de nouvelles enzymes ou bactéries. Les communautés microbiennes du rumen ont ainsi fait l'objet de nombreuses études, appliquées aux espèces locales d'intérêt (ovins (Martínez et al., 2010), bovins (Tajima et al., 1999), mais aussi renne (Sundset et al., 2009) ou même girafe (Roggenbuck et al., 2014)).

La diversité microbienne dans le rumen respecte certaines constantes, notamment en termes de Phyla représentés. La richesse en OTU ou espèces est difficile à évaluer, elle est très variables selon les publications mais est extrêmement dépendante de la profondeur de séquençage et de la qualité de la détection de chimères. Selon différents auteurs, elle peut aller jusqu'à 1000 OTUs (à 97% d'homologie). Quels que soient les espèces hôtes ou les conditions d'alimentation, on observe une forte dominance des *Firmicutes* et des *Bacteroidetes*. Près de 90% des *Firmicutes* appartiennent à la classe des *Clostridia,* représentés par les genres dominants de *Butyrivibrio, Acetivibrio,* ou *Ruminococcus*. Les *Bacteroidetes* sont eux dominés par la classe des *Bacteroidia* et le genre le plus répandu est *Prevotella.* L'analyse du microbiote du rumen bovin sur un nombre élevé d'individus, suggère l'existence d'un cœur d'une trentaine de genres partagés entre individus (Jami and Mizrahi, 2012). Mais on observe aussi des différences entre individus qui s'expliquant alors par les différences en abondances de ces genres et par les genres auxiliaires (Figure 7).

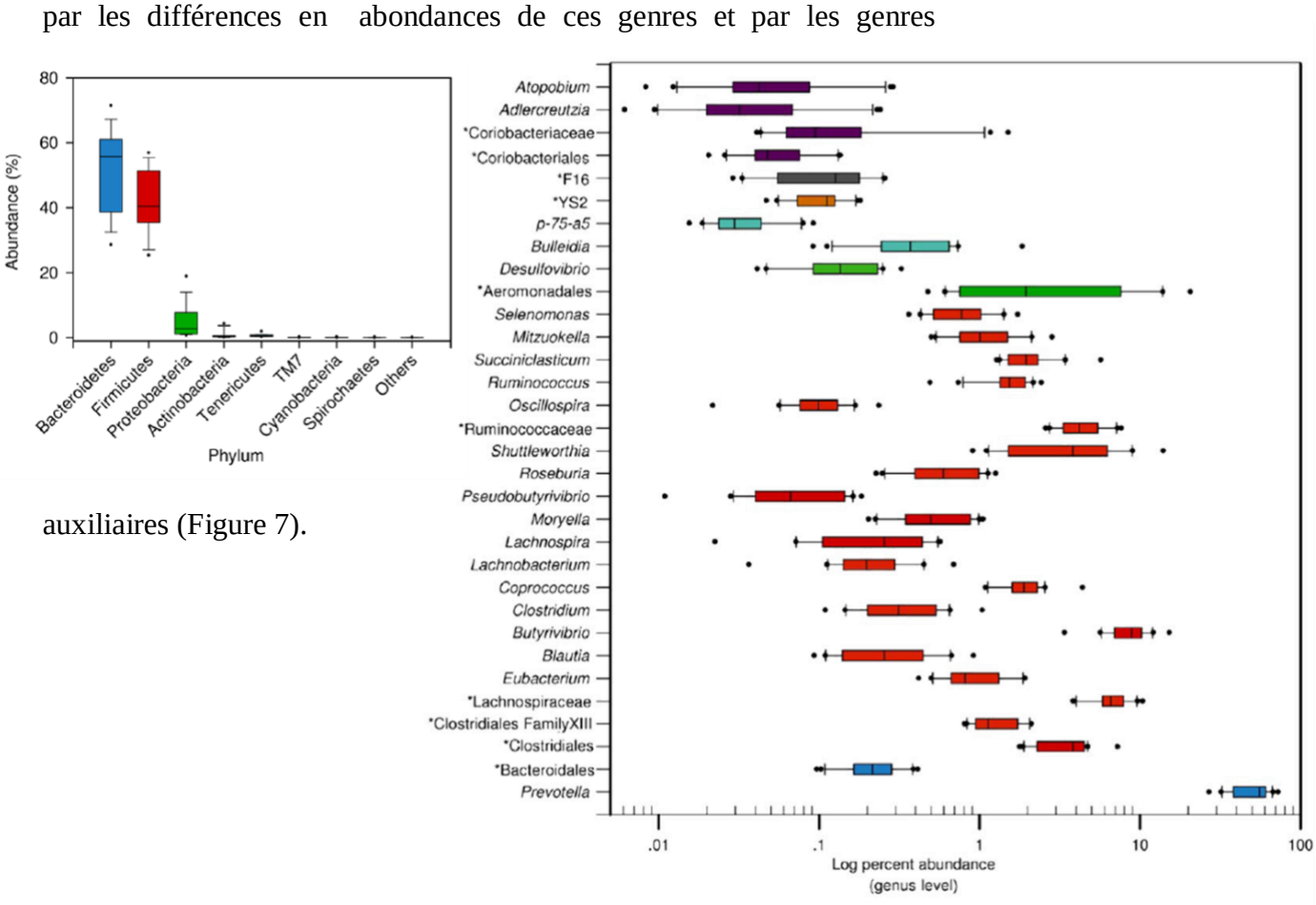


Figure 7 : Composition bactérienne du rumen bovin (d'après Jami and Mizrahi, 2012). À gauche, abondance en pourcentage au niveau du phylum. À droite, abondance des genres bactériens partagés par tous les individus. Les couleurs correspondent à leur phylum d'appartenance.

Plusieurs paramètres ont pu être identifiés comme ayant un impact sur la diversité bactérienne dans le rumen d'une même espèce hôte. Le régime alimentaire a un effet très documenté sur les populations ruminales, de nombreux travaux étudient les modifications du microbiote ruminal en réponse à un changement d'alimentation. Par exemple, le passage d'un régime à base de plantes en C4, très fibreuses, à un régime de plantes en C3, riches en protéines, cause une chute de diversité (Pitta et al., 2009). La dominance des bactéries du genre *Prevotella* et *Rikenella* (phylum *Bacteroidetes*) observée en régime C4 laisse la place à une très forte dominance du genre *Prevotella* seul (jusqu'à 56% de la communauté). Dans le contexte de l'élevage, l'utilisation de céréales peut être à l'origine d'acidoses (diminution du pH ruminal), accompagnées de changements dans les communautés bactériennes (Petri et al., 2012). Des travaux concernant des populations de rennes montrent que le régime alimentaire (naturel ou artificiel) a plus d'effet sur la diversité microbienne que l'origine géographique (Sundset et al., 2007). Cependant, si de nombreux travaux étudient la question, différencier causes et symptômes reste encore difficile et aucune généralisation du lien entre composition du régime alimentaire et diversité microbienne n'est possible, les paramètres confondants et les résultats contradictoires étant nombreux.

Néanmoins, on observe des différences de flore ruminale entre individus suivant le même régime alimentaire. Ces différences ne sont pas dues à des effets aléatoires et semblent être reliées à la biologie (pH ruminal et concentration en AGV) de l'hôte. En effet, après un échange du contenu ruminal de deux individus aux communautés bactériennes différentes, le pH et la concentration d'AGV reviennent à leur valeur de départ en quelques heures et les communautés se re-stabilisent à des compositions proches de l'initial en quelques jours (Weimer et al., 2010). Une corrélation entre des paramètres ruminaux comme la concentration en acide propionique ou la concentration en $N\text{-}NH_3$ a pu être mise en évidence (Michelland et al., 2009a). Là encore, il est impossible de discerner causes et effets. En plus, la composition du microbiote intestinal semble d'être influencée par le régime alimentaire de l'hôte mais elle semble aussi être spécifique à l'hôte, ce qui est également montré chez l'Homme (Spor et al., 2011). Enfin, pour un même individu les variations spatiales sont faibles (Li et al., 2009), le rumen étant un milieu assez homogène du fait d'un brassage élevé. On observe par contre une forte différence entre les profils bactériens observés dans le rumen et les fèces, qui sont donc une mauvaise image du fonctionnement du rumen. Par contre, certains travaux montrent une variabilité temporelle avec un intervalle d'une semaine entre prélèvements, mais pas d'effet induit par la prise d'un repas (Michelland et al., 2009a). Une modification de la flore après un

repas était pourtant établie dans des travaux plus anciens (Leedle et al., 1982), mais d'autres études plus récentes n'y voient pas non plus d'effet significatif (Li et al., 2009). Cependant, si les abondances relatives ne changent pas, il est possible que la quantité de bactéries suive des cycles de croissance/décroissance en réponse à la prise de repas.

L'étude des communautés bactériennes du rumen ne s'arrête pas au suivi de sa diversité. En effet, celle-ci peut être un bon marqueur de phénomènes dynamiques ou de différences de réponse à un traitement, mais il reste difficile de faire le lien entre phylogénie et fonction. Même s'il est maintenant possible de prédire le métagénome d'un échantillon à partir de sa diversité (Langille et al., 2013), ces prédictions peuvent s'appliquer à des environnements à la flore très bien caractérisée. Dans le cas où ceux-ci sont très peu caractérisés et composés de bactéries inconnues ou au génome non disponible, une telle prédiction perd en précision et une approche métagénomique classique est indispensable à l'étude des fonctions réalisables par une communauté bactérienne.

Une approche métagénomique a permis d'étudier les différences de communautés bactériennes impliquées dans les variations individuelles de production de méthane observées chez les bovins ; la production de méthane pouvant aller de 7,6 à 32,4 g/kg de matière ingérée (Wallace et al., 2015). Les individus très méthanogènes présentent une flore 2,5 fois plus riche en archées, notamment *Methanobrevibacter*, mais les indices de diversité ne montrent pas de changements significatifs entre les deux groupes étudiés (faibles et forts producteurs de méthane). Cependant, certaines enzymes clés de la production d'acides gras volatils, comme les acétates kinases, les pyruvate-formate-lyases, et d'autres sont différemment présentes entre groupes, favorisant une voie de formation d'acétylCoA à partir de pyruvate chez les faibles producteurs de méthane. Cette voie est par ailleurs la voie par laquelle le pyruvate est converti en acétate chez une souche de *Succinivibrionaceae*, une famille dont l'abondance est plus forte chez les faibles producteurs de méthane. L'approche métagénomique permet donc de détecter des différences d'abondances en gènes, corrélées à des variations de diversité, et à expliquer des différences de production de méthane par le rumen.

L'analyse métagénomique du microbiome bovin adhéré au substrat a montré que cette flore possède un grand nombre de gènes liés à l'hydrolyse des chaînes latérales de l'hémicellulose, mais relativement peu de gènes de l'hydrolyse des chaînes principales ou de la cellulose (Brulc et al., 2009). De plus, peu d'entre ces gènes possèdent des domaines de liaison à la cellulose. Ces résultats étant contradictoires avec d'autres travaux, les auteurs

proposent que leurs données permettent de caractériser la flore qui adhère initialement au substrat, capable d'une hydrolyse rapide de la fraction facilement hydrolysable, et que celle-ci soit ensuite remplacée dans un second temps par une flore capable de dégrader les chaînes principales d'hémicellulose et de cellulose.

Enfin, avec une très grande profondeur de séquençage, la métagénomique permet de reconstruire le génome de bactéries constituant la communauté, même si celles-ci ne sont pas cultivées ou cultivables. En effet, dans l'étude de Hess et al. (2011), moins d'un pourcent des assemblages de lectures présentent une forte identité avec des génomes de référence, mais il est possible de reconstruire les génomes constituant du métagénome en utilisant les données de couvertures et de pourcentage en GC. Chaque génome ainsi reconstitué peut de plus être associé à son gène codant pour l'ARNr 16S, ce qui donne accès à une interprétation fonctionnelle des données de diversité.

II.3.2 Études en systèmes artificiels : bioréacteurs

L'utilisation de communautés microbiennes pour leur capacité à dégrader la lignocellulose en bioréacteurs a déjà été étudiée dans différents contextes, qui ne se limitent pas à la plateforme de carboxylates. Les conditions utilisées peuvent parfois être très différentes, notamment lorsque les cultures sont faites en présence d'oxygène ou à haute température. Cependant, même si les différences de modes opératoires, de sources d'inoculum ou de substrat rendent les comparaisons directes difficiles, les résultats de ces travaux apportent déjà des éléments d'analyse et de compréhension des communautés microbiennes lignocellulolytiques en réacteurs.

.II.3.2.1 Enrichissements

De nombreuses études présentent les résultats d'enrichissements de microorganismes lignocellulolytiques en réacteurs utilisant un substrat lignocellulosique comme seule source de carbone. L'enrichissement est une stratégie utilisée pour plusieurs raisons : en augmentant la proportion d'organismes actifs, elle permet d'augmenter les performances d'un inoculum, mais aussi de le simplifier et de faciliter l'étude de son fonctionnement. Enfin, dans le cadre d'une utilisation directe de communautés microbiennes (cas de la plateforme des carboxylates

ou de méthanisation), elle permet la stabilisation d'un consortium microbien à partir d'un inoculum non acclimaté aux conditions de culture. L'enrichissement consiste le plus souvent en un enrichissement par sous-cultures, ou repiquage : une fraction de la fin d'un réacteur conduit en mode batch est conservée, et sert d'inoculum à la culture suivante. L'inoculum de départ est ainsi dilué à chaque cycle, et seuls les microorganismes ayant un taux de croissance suffisant dans les conditions imposées se maintiennent dans la culture. Ce type de sélection peut en théorie être appliqué en réacteurs continus plutôt qu'en batch successifs, mais l'utilisation d'un substrat solide complique alors la conduite du procédé.

Plusieurs expériences d'enrichissement ont été menées sur des nombres élevés de cycles. Après 45 cycles, Peng et al. (2010) ont obtenu une communauté très simplifiée, de 19 espèces, présentant des niveaux d'activité xylanase très élevés. Cette communauté était capable de dégrader 50% de la paille de maïs utilisée comme substrat, utilisant préférentiellement la fraction hémicellulosique (80%) puis la fraction cellulosique (30%). Leur analyse montre que la communauté était stable sur les 35 derniers cycles de culture étudiés ; la stabilisation à partir de l'inoculum provenant du sol amendé en compost n'ayant requis qu'une dizaine de cycles. Les bactéries identifiées sont des *Bacteroidetes* (*Bacteroides, Dysgonomonas* et inconnus), des *Clostridiales* (8 clones de *Clostridium* et quelques autres moins caractérisés), et des *Proteobacteria* variés, aérobies connus pour la plupart (*Pseudomonas, Alcaligenes, Escherichia*). La diversité de l'inoculum de départ n'a malheureusement pas été caractérisée. Néanmoins, en se rapportant à la diversité bactérienne des sols, généralement dominée par les phyla *Proteobacteria, Acidobacteria* et *Actinobacteria* (Janssen, 2006), il est possible de remarquer que les conditions de culture ont fortement modifié la communauté initiale. Le type de profil de la communauté enrichie, dominé par des *Clostridiales* (ou plus largement *Firmicutes*) et des *Bacteroidetes* est assez habituel dans de nombreux écosystèmes digestifs, est régulièrement décrit après des procédures d'enrichissement en réacteur, que les conditions soient aérobies thermophiles (Eichorst et al., 2014), anaérobies thermophiles (Wongwilaiwalin et al., 2010; Ji et al., 2012), aérobies mésophiles (Feng et al., 2011) ou anaérobies mésophiles (Gao et al., 2013).

L'évolution de la communauté pendant un processus d'enrichissement a également été étudié par Feng et al. (2011). Après 40 cycles de repiquage, la diversité bactérienne montre une faible évolution entre les cycles 10 et 40 (les résultats pour les cycles précédents ne sont pas présentés). Cependant, si les espèces enrichies restent stables, les abondances relatives sont encore variables et on observe un changement d'espèces dominantes. Par ailleurs, les

auteurs ne constatent aucune amélioration des capacités de dégradation au cours de l'enrichissement, leur inoculum étant capable de dégrader environ 85% de papier filtre, ou 50% de résidus de maïs après 8 jours de culture, et ce dès les premiers cycles. Dans cette étude, l'enrichissement abouti donc à la stabilisation d'une communauté active pour la dégradation de lignocellulose, mais n'augmente pas pour autant ses performances. Il est intéressant de noter que dans cette étude, 4 inocula différents ont été utilisés mais que 2 d'entre eux ont perdu toute activité de dégradation dans les premiers cycles de culture sur lignocellulose. Seul l'inoculum le plus performant a été conservé pour la poursuite de l'enrichissement. Les auteurs mettent en avant les caractéristiques du milieu de culture comme explication ; les inocula présentant les meilleures capacités de dégradation étant ceux pour lesquels les conditions de pH initiales étaient proches des conditions de culture.

L'importance des conditions de culture est également mise en évidence par les travaux de Lü et al. (2012) sur l'effet de la température. En partant de la même source d'inoculum (du compost) mais en procédant à deux enrichissements indépendants, l'un à 50°C et le second à 60°C. Les auteurs obtiennent deux communautés aux capacités de dégradations différentes de, respectivement, 46% et 60% de dégradation de paille de blé. Cependant, en inversant les températures d'incubation pendant 8 nouveaux cycles, les capacités de dégradation s'en trouvent inversées également. Les deux communautés présentent beaucoup d'espèces communes, c'est surtout leurs abondances relatives qui sont changées entre les communautés enrichies. Quand les températures de culture sont modifiées, les communautés montrent une certaine résilience, mais ont tendance à se ressembler davantage entre températures communes. Ceci suggère l'adaptabilité des communautés microbiennes à des variations des conditions de culture. Si l'expérience avait était prolongée sur plus de 8 cycles, on aurait probablement observée une convergence.

Enfin, il faut également garder à l'esprit que si les conditions de culture sont assez différentes de celles de la source d'inoculum, la forte pression de sélection qui en résulte peut amener entrainé des importants effets aléatoires. Ainsi, dans l'étude de Eichorst et al. (2013), un enrichissement est réalisé en triplicat sur deux substrats différents, les conditions de culture et sources d'inoculum étant identiques. Les auteurs observent des différences en composition en *Bacteroidetes* dès le premier cycle d'enrichissement sur cellulose. Ces différences s'amplifient dans les cycles suivants pour aboutir à des communautés à dominance soit *Bacteroidetes*, soit *Firmicutes* (>70%). Ces différences ne sont pas nécessairement maintenues d'un cycle à l'autre et peuvent présenter une certaine oscillation. Au niveau

OTU, l'enrichissement sur cellulose présente une faible richesse, ce qui peut rendre d'autant plus forts des effets aléatoires. Cependant, l'expérience n'a été conduite que sur une durée de 4 cycles, on ignore donc si les différents réplicats auraient fini par se stabiliser à des compositions similaires les unes aux autres.

Par ailleurs, dans des études sur l'effets de la source d'inoculum et du substrat utilisé, Simmons et al. (2014) présentent des résultats où les triplicats ont des comportements répétables. À partir de deux différents composts utilisés comme inoculum pour deux substrats différents, eucalyptus et « switchgrass » (une Poacée sauvage américaine), les communautés initiales et finales après 15 jours d'incubation montrent une bien meilleure répétabilité entre triplicats, même si certains points sont assez distants entre eux (Figure 8). Dans cette étude, le substrat semble avoir une influence beaucoup plus forte que la source d'inoculum sur la structure finale des communautés.

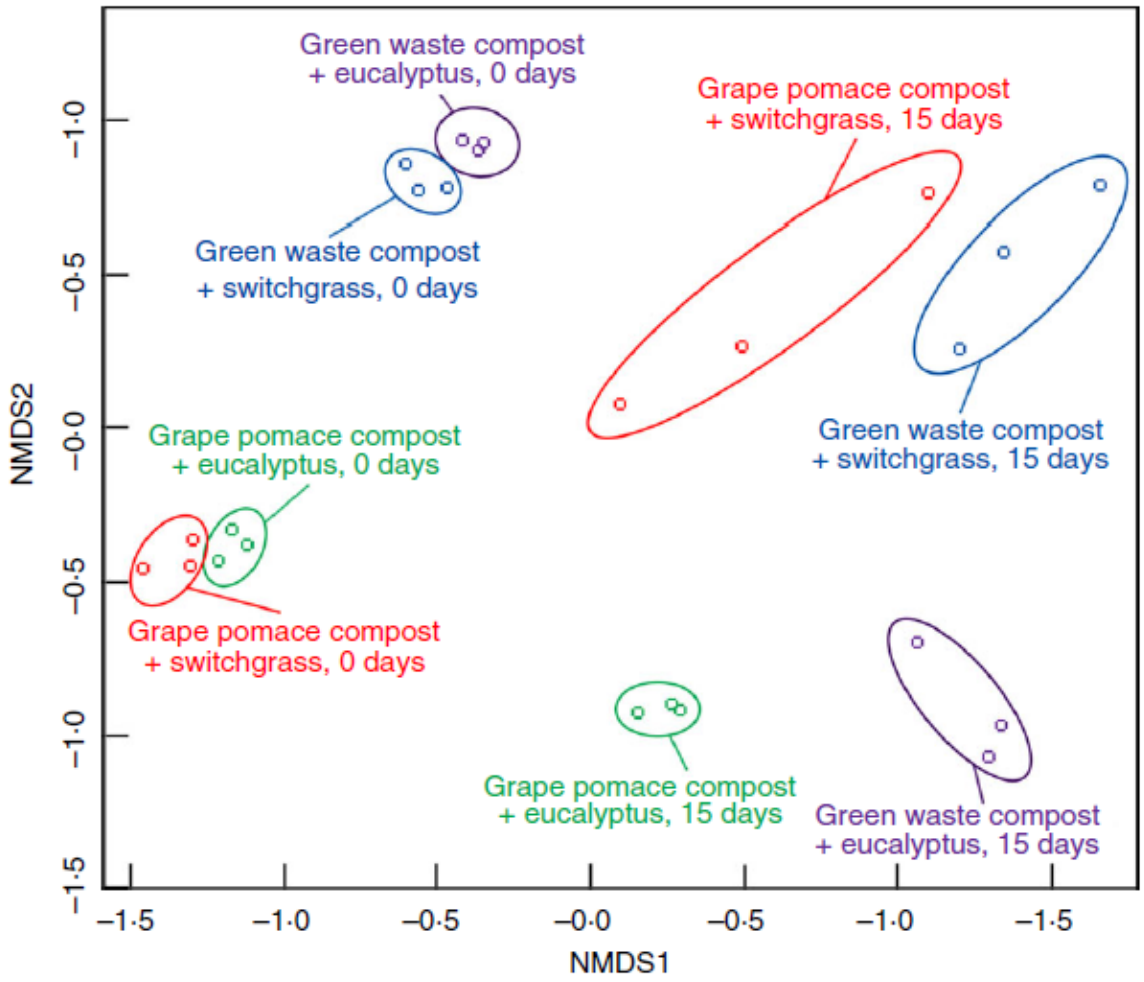


Figure 8 : Graphe NMDS basé sur les valeurs de dissimilarité de Bray-Curtis (d'après (Simmons et al., 2014)) 4 combinaisons différentes inoculum+substrat ont été étudiées. Les points initiaux sont regroupés par source d'inoculum, « green waste compost » et « grape pomace compost», tandis que les communautés finales sont davantage influencées par le substrat (eucalyptus en bas à droite, switchgrass en haut à droite).

.II.3.2.2 Évolution de la diversité au cours de la dégradation

Pendant un cycle de dégradation, le suivi au cours du temps des différents paramètres de culture permet d'en apprendre davantage sur le fonctionnement d'un inoculum. Si de nombreuses études ne s'attachent qu'à étudier des paramètres de fin de cycle, quelques-unes s'intéressent tout de même à leur dynamique en cours de cycle. Hui et al. (2013) ont étudié les dynamiques de dégradation d'un consortium enrichi sur 3 substrats (paille de blé, de riz, et résidus de maïs) en effectuant des mesures tous les 3 jours d'un cycle de 12 jours. Les résultats varient en fonction du substrat, mais montrent des variations de pH assez fortes (diminution initiale, suivie d'une augmentation lente), ainsi que des pics d'activité xylanase excrétée autour du jour 6 (paille de blé et maïs). La dégradation des pailles de blé et de maïs atteint un plateau dès le jour 6, tandis que celle des résidus de maïs se poursuit à un rythme plus lent, en étant associée à des niveaux d'activité xylanase toujours croissants. La dynamique de production d'acides volatiles suit logiquement le même profil, la majeure partie d'entre eux étant produits pendant les 6 premiers jours du cycle. L'analyse de la dynamique des communautés bactériennes par DGGE montre elle une forte stabilité de la communauté. Les 8 bandes visibles sont décrites comme stables au cours du temps, même si on peut voir quelques bandes dont l'intensité (faible relativement aux bandes dominantes) augmente légèrement au cours du temps ou présentent des pics à 3 ou 6 jours, correspondant à *Bacteroidales* et *Clostridium*. Ce résultat est assez surprenant au vu des résultats qui montrent le très fort effet que peut avoir le pH sur les communautés bactériennes, mais l'échelle de temps est peut-être trop faible pour avoir un effet sur l'abondance des populations qui serait visible en termes de quantité d'ADNr 16S. Lü et al. (2012) présentent des résultats de dynamique avec de nombreux points de prélèvement au cours d'un cycle de 40 jours, et montrent une augmentation continue du pH, un pic d'activité CMCase (représentative de l'activité cellulase) autour du jour 16 et un pic de biomasse aux alentours du jour 7. La température choisie a un effet sur les dynamiques, qui sont plus précoces et aux valeurs plus extrêmes à 60°C qu'à 50°C. Malheureusement, la communauté bactérienne n'est pas suivie au cours du temps.

Zhao et collègues ont caractérisé la dynamique de dégradation de leur inoculum (issu de compost) au cours de cycle de 12 jours de dégradation de paille de riz (Zhao et al., 2014a). La dégradation des différents composants de la lignocellulose atteint un plateau après 9 jours, ce qui correspond également au maximum d'activité cellulase et xylanase, ainsi qu'au maximum de diversité (estimée par DGGE). Étonnamment, un pic de biomasse très clair,

mesuré par qPCR du gène de l'ARNr 16S est observé au jour 1. Quatre des bandes correspondantes en DGGE sont de forte intensité uniquement aux jours 1 et 3, et correspondent toutes à des *Clostridium (*sauf une *Clostridiales* indéterminée). À l'inverse, d'autres bandes n'apparaissent que plus tardivement, correspondant à *Bacillus* ou des bactéries moins bien identifiées (*Clostridiaceae)*. Contrairement aux résultats de Hui et al. (2013), on observe donc ici une dynamique assez forte de la communauté bactérienne au cours de la dégradation.

Enfin, dans leur suivi de communauté en cours d'enrichissement, Eichorst et al. (2013) ont réalisé une caractérisation de la dynamique de la communauté lors du dernier cycle d'enrichissement. Les trois réplicats d'enrichissement présentent une forte variabilité en termes de composition, déjà discutée précédemment, mais le suivi des populations bactériennes pour un même réplicat au cours du temps montre de fortes variations. Les *Bacteroidetes*, représentés par un OTU majoritaire, présentent de fortes variations, oscillantes entre les 4 temps mesurés, et les *Firmicutes*, plus diverses et qui représentent la principale autre population, oscillent en conséquence. Certains changements sont drastiques : pour un des réplicats, une disparition totale des OTUs *Firmicutes* est observée au jour 6 en faveur d'un unique OTU au métabolisme improbable (aérobie marin halophile), qui disparait à son tour au jour 10, remplacé par un profil en *Firmicutes* proche de celui du jour 3. Ces changements sont peu plausibles et pourraient correspondre à des problèmes d'amplification ou de séquençage ; ainsi, il est nécessaire de conserver un regard critique sur ces résultats.

.II.3.2.3 Effet des prétraitements

Les prétraitements de la biomasse lignocellulosique sont une étape très souvent incontournable des procédés de valorisation, comme cela a été discuté dans la partie dédiée. S'ils ont le plus souvent été étudiés en lien avec les modifications structurelles du substrat et des paramètres qui y sont liés (porosité, gonflement, cristallinité…) ainsi qu'à la réponse à des traitements enzymatiques, quelques travaux s'intéressent à l'impact du prétraitement du substrat sur les communautés microbiennes qui le dégradent. Le prétraitement est supposé augmenter la digestibilité, notamment en améliorant l'accessibilité de la cellulose et de l'hémicellulose au sein de la structure complexe de la lignocellulose (Rollin et al., 2011). Le substrat prétraité, modifié, peut donc être vu comme un substrat différent, ou comme une

modification des conditions de culture, tous deux étant capables d'avoir un impact sur les communautés lignocellulolytiques, comme discuté dans les deux parties précédentes. Une partie des articles s'intéressant à la digestion par des communautés microbiennes de substrats prétraités ne comporte malheureusement pas d'étude de l'impact sur la diversité. Zhao et al. (2014b) par exemple teste les conditions de prétraitement thermochimique (en présence d'acide acétique) qui ont le plus d'effet sur la digestion par un inoculum (boue de digesteur) mais ne s'intéresse pas aux variations de diversité en lien avec l'effet de ces prétraitements sur le substrat. Wen et al. (2015) testent la réponse de 3 inocula à un prétraitement en suivant les activités enzymatiques, la production de biomasse et la dégradation de chaque composant de la lignocellulose mais ne réalisent pas de suivi des communautés bactériennes associées. Enfin, Guo et al. (2011) étudient l'impact de 5 prétraitements différents sur la production d'acides gras volatils par un consortium microbien MC1, mais là encore ne caractérisent que des paramètres de la dégradation (évolution de la dégradation, du pH et des concentrations en différents acides gras volatils) sans s'intéresser à la diversité.

Dans une étude récente Eichorst et al. (2014) s'intéressent à l'adaptation d'une communauté (thermophile aérobie) à des substrats modifiés par deux prétraitements, AFEX et IL. L'évolution de la communauté (issue de compost) au cours d'un enrichissement par repiquage, en triplicat, a été caractérisée par séquençage du gène de l'ARNr 16S. Leurs résultats montrent que les prétraitements n'ont pas le même impact sur la diversité. Comparé à l'enrichissement sur substrat non prétraité, AFEX amène à une stabilisation de la diversité (richesse, Shannon et Simpson) à des valeurs plus élevées alors qu'IL produit une communauté à la diversité plus faible. Partant d'un inoculum composé majoritairement de *Firmicutes* et *Actinobacteria*, la communauté témoin se stabilise à 40% *Bacteroidetes*, 25% *Firmicutes* et 25% *Thermi*. La communauté AFEX est plus riche en *Firmicutes* au détriment des *Thermi* alors que la communauté IL est largement dominée par les *Bacteroidetes* (>50%) avec une quasi absence du phylum *Thermi* (phylum des *Deinococcus*). Le prétraitement utilisé ici, très agressif, a un effet non négligeable sur les communautés sélectionnées, ce qui au niveau microbien se traduit par la sélection de bactéries aux potentialités de dégradation du substrat très différentes. Au niveau OTU, on observe cependant peu de spécificité stricte, et les OTUs majoritaires dans une condition de prétraitement sont, à deux exceptions près, présents mais à une abondance plus faible, dans les autres conditions. Ces résultats indiquent donc que le prétraitement d'un substrat peut avoir une influence non négligeable sur la composition des communautés qui le dégradent. Ceci peut avoir à son tour un impact sur les

performances de dégradation, mais également sur les proportions d'acides gras volatils produits.

Dans le contexte d'un procédé à forte teneur en solide, Reddy et al. (2012) ont testé l'effet la présence d'un liquide ionique (le 1-éthyl-3-méthylimidazolium acétate, permettant une meilleure solubilisation de la biomasse) sur l'enrichissement des communautés. L'enrichissement en l'absence de liquide ionique fait passer la communauté d'une forte diversité (inoculum avec un indice de Shannon de 5,81 et une richesse de 1000) à une diversité beaucoup plus faible, stabilisée autour d'un Shannon de 1,4 pour ~100 de richesse. En présence de liquide ionique, cette diversité est encore plus faible, avec un Shannon de 0,68 et environ 80 de richesse en fin d'enrichissement. Un tel prétraitement peut donc avoir un effet très négatif sur la diversité, en conduisant à l'élimination de certains OTUs et en favorisant l'extrême dominance d'autres (jusqu'à 85,5% d'abondance pour l'OTU dominant en présence de liquide ionique). Cependant, maintenir une communauté active en présence de liquide ionique était déjà un succès pour les auteurs, qui lors de précédents travaux étaient incapables d'obtenir une activité dans de telles conditions.

En résumé, les communautés microbiennes permettant la dégradation de lignocellulose en bioréacteurs commencent à être assez bien décrites. Dans la plupart des cas, quelle que soient les sources d'inoculum, elles sont dominées par les phylums *Firmicutes* et *Bacteroidetes,* et ressemblent donc plus aux communautés digestives qu'aux communautés du sol. De nombreux travaux montrent que les conditions de culture et le choix du substrat ont un impact plus fort sur la communauté finale que l'inoculum de départ. DeAngelis et al. (2012) ont ainsi montré que l'ajout de minéraux ou métaux, impliqués dans les réactions de transferts d'électrons pendant la fermentation, pouvaient avoir un rôle primordial sur les bactéries sélectionnées. La culture en milieu souvent minimum et à forte teneur en eau pourrait donc être une force de sélection favorisant les phylums cités, que la température soit élevée ou faible, et en conditions aérobies ou anaérobies. La source d'inoculum a cependant son importance et son empreinte est conservée dans les communautés obtenues, mais l'influence des conditions de culture semble être la plus forte.

III. Méthodes et outils d'étude des communautés bactériennes

Parmi les micro-organismes présentés précédemment, seuls quelques-uns ont pu être cultivés, mais présentent alors des capacités hydrolytiques assez faibles, ce qui révèle l'importance des interactions entre partenaires au sein de communautés microbiennes. Ces communautés microbiennes sont naturellement ubiquistes dans l'environnement et jouent des rôles clés dans des processus majeurs comme les cycles du carbone ou de l'azote. Leurs rôles écologiques sont nombreux, mais leur action est souvent complexe, faisant intervenir différents partenaires en synergie ou en successions spatiales et temporelles. L'écologie microbienne, qui étudie les interactions entre micro-organismes et s'intéresse à tous les phénomènes qui y sont liés, est un domaine de recherche immense du fait de l'omniprésence des communautés microbiennes. Elle intervient dans des domaines allant de la santé humaine à la biologie marine, en passant par l'agronomie.

Si la compréhension du fonctionnement des écosystèmes microbiens en est encore à ses débuts et que les possibilités en recherche fondamentale sont encore innombrables, le potentiel des communautés microbiennes pour des applications en biotechnologie est également immense. Leurs génomes sont une source intarissable de nouvelles enzymes et voies métaboliques pour des applications diverses. Elles sont à la base de procédés comme l'épuration de l'eau ou des fermentations alimentaires, où elles ont été utilisées avant même de comprendre leur fonctionnement. En utilisation directe, les communautés microbiennes permettraient de réduire les coûts liés aux besoins de stérilité indispensable pour des cultures pures, tout en limitant le recourt à des organismes génétiquement modifiés dû leur plus grand potentiel génétique. Cependant, l'utilisation maîtrisée de communautés microbiennes nécessite une bonne compréhension de leur fonctionnement.

L'étude de la composition des communautés microbiennes a longtemps reposé sur l'isolement de différentes souches puis leur identification individuelle par des méthodes classiques (e.g. morphologie, utilisation de divers substrats). La découverte du gène codant pour l'ARNr 16S a permis de développer l'analyse des communautés microbiennes (Woese and Fox, 1977). L'ADNr16S présente en effet des régions constantes permettant son amplification simultanée chez différents espèces, et des régions variables entre espèces permettant de les discerner. L'amplification d'échantillons environnementaux permet d'étudier des communautés bactériennes en suivant la diversité des profils de migration de gènes ou marqueurs universels (Giovannoni et al., 1990). Les progrès des méthodes de

séquençage ont ensuite permis de séquencer l'ensemble des séquences d'ADNr 16S amplifiées, et d'obtenir des profils d'abondance relative.

III.1. Gènes utilisés en écologie moléculaire

III.1.1 ARNr 16S

D'abord proposé comme horloge moléculaire, le gène de l'ARNr 16S est historiquement le premier à avoir été utilisé pour établir la phylogénie d'organismes non cultivés ou non connus (Zuckerkandl and Pauling, 1965). Depuis, il est devenu un des marqueurs les plus utilisés en écologie microbienne. Impliqué dans la synthèse des protéines, il est présent chez tous les procaryotes, et présente un équivalent Archée mais également eucaryote (l'ARNr 18S) dont certaines portions sont homologues avec l'ARNr 16S. Il présente toutes les caractéristiques attendues d'un marqueur moléculaire : des portions conservées, cibles idéales d'amorces de PCR, encadrent des régions hypervariables qui permettent de différencier et d'assigner une séquence à une espèce à l'aide de banques de données. Son défaut est d'être présent en multiples copies dans les génomes, ces copies n'étant pas systématiquement strictement identiques. Ces copies variantes peuvent alors être à l'origine de plusieurs 'ribotypes' pour une seule bactérie (Case et al., 2007). L'ARNr 16S, long d'environ 1500 paires de base, comporte 9 régions hypervariables, nommées V1 à V9. Celles-ci sont de longueur variable, de quarante paires de base pour les plus petites à une centaine pour les plus grandes, V3 et V9. Les régions hypervariables sont polymorphiques en termes de substitutions, mais aussi d'insertion et délétions qui conservent la structure secondaire de l'ARN (Figure 9).

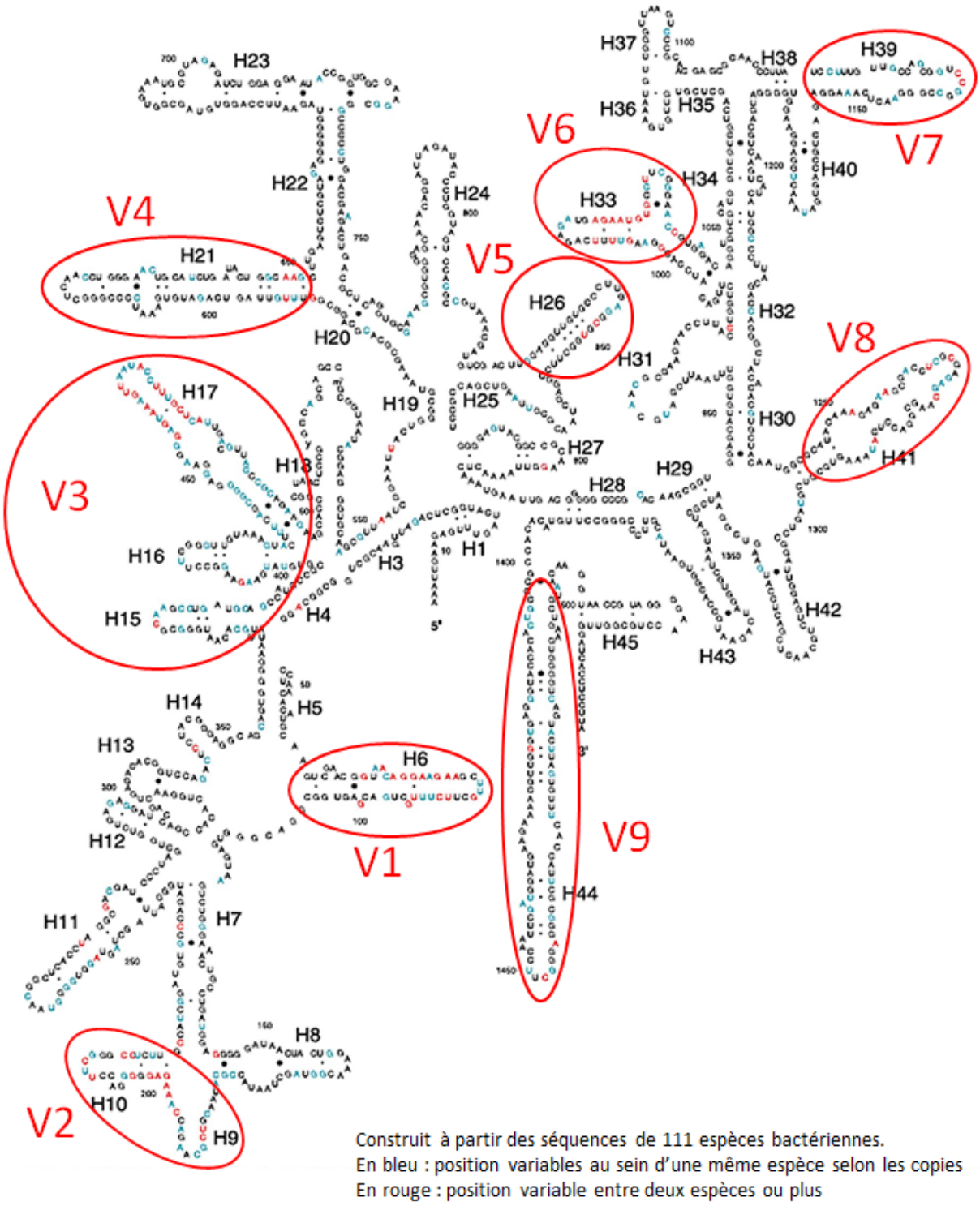


Figure 9 : Structure secondaire de l'ARNr 16S (d'après Case et al. 2007).
La séquence présentée ici est celle d'E. coli, les bases en couleur sont les positions variables au sein de l'espèce (bleu) ou entre espèces (rouge).

Il n'est pas nécessaire de disposer de toute la séquence pour que le gène ARNr 16S soit informatif, de petits fragments de 100 ou 200 paires de bases peuvent être suffisants (Liu

et al., 2007). Cependant, le choix de la région à séquencer peut être critique. Dans certain taxon, le polymorphisme peut être concentré dans des régions particulières, et de ce fait, certaines régions sont alors plus résolutives que d'autres (Kumar et al., 2011). Les couples d'amorces choisis peuvent être à l'origine de biais d'amplification entre genres bactériens, et de manière plus générale de nombreux travaux montrent des différences assez fortes entre les résultats obtenus sur des échantillons identiques, mais en ciblant des régions différentes (Winsley et al., 2012). Il faut donc être conscient lors de l'étude de communautés bactériennes, que les résultats produits ne sont pas absolus mais sont une image de l'échantillon qui peut contenir certains biais. En plus des biais propres aux techniques de séquençage ou d'analyse, l'extraction de l'ADN (des biais de rendement peuvent exister entre espèces), l'amplification (biais d'amplification en fonction des taxons, différents selon les régions hypervariables), et le nombre de copies peuvent être à l'origine de différences entre les résultats obtenus par différentes méthodes et la réalité.

III.1.2 Autres marqueurs

À la différence de l'ARNr 16S, de nombreux autres marqueurs ne présentent pas l'inconvénient du nombre variable de copies et sont quasiment aussi universels. Ainsi, le gène codant pour la sous-unité β de l'ARN polymérase bactérienne, rpoB, présent systématiquement en une seule copie, a notamment été présenté comme un marqueur permettant une meilleure résolution que l'ARNr 16S (Case et al., 2007). D'autres 'gènes de ménage' peuvent également jouer ce rôle, tels que gyrB, codant pour la sous-unité de la DNA gyrase, ou dnaK, codant pour une protéine chaperonne régulatrice du stress thermique (Liu et al., 2012b). Ces marqueurs, souvent présentés comme bien meilleurs que l'ARNr 16S, restent cependant moins utilisés principalement à cause du manque d'information dans les bases de données correspondantes comparé à la richesse des bases de données ribosomiques.

Certains travaux s'intéressent à des gènes fonctionnels qui leur permettent d'étudier uniquement des populations d'intérêt. C'est le cas par exemple des gènes impliqués dans l'oxydation de l'ammonium amoA ou amoB, qui ne sont pas universels, mais qui ne sont présents que chez les archées (amoA) ou bactéries (amoB) capables d'oxider l'ammonium. D'autres marqueurs, tels que les gènes codants des di-oxygénases permettent d'identifier les microorganismes impliqués dans la dégradation des hydrocarbures, et on peut également cibler une famille de glycoside hydrolases. L'utilisation d'un marqueur moléculaire

fonctionnel est une façon différente d'aborder l'écologie moléculaire, mais peut être complémentaire d'une approche de diversité globale plus classique (Xu et al., 2012).

Les marqueurs moléculaires ciblant des gènes codants ont cependant un inconvénient majeur dont le ARNr 16S est affranchi : la 3^{e} position de chaque codon étant très souvent muette, les mutations silencieuses s'y accumulent fortement et sur un temps d'évolution assez long, vont jusqu'à saturation (chaque nucléotide étant équiprobable, au taux de GC près). Si une variabilité élevée est un avantage dans la région discriminante d'un marqueur, elle pose un problème pour le dessin d'amorces universelles. Ainsi, les biais d'amplification (liés à la bonne hybridation-élongation en cours de PCR) sont donc potentiellement plus forts avec les marqueurs codants qu'avec le gène de l'ARNr 16S, non codant mais multi-copie.

Un autre marqueur moléculaire non codant est utilisé, l'ITS (pour Internal transcribed spacer). Celui-ci correspond à une insertion entre les gènes codant pour les différents ARNr. Étant non fonctionnelle, cette région est extrêmement variable même entre espèce très proches, et est encadrée par les régions conservées des ARNr. Le marqueur ITS est peu utilisé chez les procaryotes, mais est couramment utilisé en biologie végétale (résolution de la phylogénie des Asteraceae (Baldwin, 1992)) et chez les champignons, où il est le marqueur moléculaire le plus utilisé, à la fois en écologie moléculaire, en identification et même en diagnostic clinique (Chen et al., 2001). Concernant l'écologie fongique, cette habitude s'appuie sur des études comparatives de marqueurs. Le Fungal Barcoding Consortium a comparé les six régions les plus utilisées en écologie (trois régions ribosomiques et trois gènes codants) afin de choisir le meilleur marqueur moléculaire pour les champignons (Schoch et al., 2012). Les gènes codants présentent les meilleurs taux d'identification, mais des efficacités de PCR et un succès de séquençage plus faibles ; ils ont donc été éliminés. Parmi les régions ribosomiques, l'ITS présente les meilleurs taux d'identification pour une très large partie des taxons étudiés, ainsi qu'un décrochement permettant de discerner variation intra- et inter-spécifique. La petite sous-unité ribosomique présente une faible résolution, et la grande sous-unité, si elle a une résolution plus forte que l'ITS pour certains groupes (lignées très anciennes et Ascomycètes), a une résolution plus faible pour tous les autres.

III.1.3 Banques de données

Si tous les marqueurs décrits précédemment peuvent être utilisés avec des techniques sans séquençage (voir partie suivante), leur emploi dans un but d'identification repose sur la qualité des banques de données. Les séquences présentes dans les banques servent de référence pour l'assignation taxonomique, et la richesse et la précision des banques est donc déterminante pour une bonne identification. Les marqueurs moléculaires peu courants ont donc pour inconvénient supplémentaire de ne souvent disposer que de banques de petite taille. Les marqueurs très courants, comme l'ARNr 16S disposent parfois de plusieurs banques de données différentes, qui ont toutes leurs avantages et inconvénients.

Silva est une des banques des séquences d'ARNr 16S les plus riches qui, de plus, est mises à jour très fréquemment (Quast et al., 2013). Dans ses dernières versions, une étape de non-redondance a été appliquée pour contrer une richesse en trop forte augmentation qui la rendait coûteuse en ressources. Un critère de 99% d'identité est maintenant appliqué à l'aide de l'outil UCLUST afin d'éliminer les séquences très similaires. Le nombre d'entrées est fortement réduit mais la représentativité est ainsi conservée. La dernière version, la 123, contient 597 607 séquences, contre près de 5 millions au total et 1,7 millions de séquences uniques. Un ensemble d'outils (ARB) est distribué par Silva.

La base de données LTP (All-Species Living Tree Project) est une sous-partie 'nettoyée' de Silva et contient un nombre beaucoup plus faible de séquences (11 900 dans sa dernière version)(Munoz et al., 2011). Les séquences présentes correspondent uniquement à des souches types d'espèces bien classifiées d'Archées et de Bactéries. Son avantage est un poids très faible, ce qui a un impact important sur les temps de calcul et une 'propreté' inégalée avec uniquement des séquences vérifiées définies jusqu'à l'espèce voire la souche. Cependant, pour des études environnementales ou concernant des communautés faiblement caractérisées et peu connues, elle est peu informative. Associée à des méthodes k-mer, une représentativité trop faible peut également amener des erreurs d'assignation.

RDP est à la fois une banque de données et un ensemble d'outils, dont une méthode d'assignation par k-mer (Cole et al., 2014). Cette banque est pour l'instant non nettoyée et regroupe plus de 3 millions de séquences ; ceci la rend lourde à manipuler. D'autre part, RDP met à disposition une infrastructure d'assignation en ligne, et sa méthode est une des plus reconnues, elle est ainsi l'une des plus utilisées.

Greengenes est une base réputée pour sa 'propreté', avec des séquences vérifiées manuellement et ne contient aucune chimère, un problème courant dans d'autres bases de données (DeSantis et al., 2006). Avec un peu plus d'un million de séquences, elle contient moins de séquences que Silva ou RDP mais sa représentativité pour tous les phyla est très bonne. En plus, son format est compatible avec l'outil ARB de Silva. Cependant, les mises à jour sont assez rares, la dernière remontant à 2013. Elle est reste assez utilisée.

Enfin, de nombreux travaux utilisent des bases de données spécifiques à leur objet d'étude, soit construites *de novo,* soit enrichies à partir d'une base existante. C'est le cas de DictDB, une banque dédiée à l'étude des flores digestives d'insectes (Mikaelyan et al., 2015b). Basée sur Silva, DictDB incorpore les données de clones obtenus à partir d'études de microbiotes intestinaux d'insectes, pour lesquels la résolution et la précision en sont ainsi augmentées. L'avantage de ce type d'approche est de pouvoir inclure des séquences déjà trouvées dans d'autres échantillons (même si elles sont inconnues), tout en pouvant se servir d'une base de données plus légère, contenant que des séquences pertinentes pour l'environnement étudié.

III.2. Techniques de suivi

Une fois un marqueur moléculaire choisi, il existe différentes techniques pour l'étude des communautés bactériennes, plus ou moins récentes et impliquant ou non du séquençage.

III.2.1 Les techniques sans séquençage

Avec une électrophorèse classique, des doubles brins d'ADN peuvent être séparés en fonction de leur taille ; avec certains outils comme l'électrophorèse capillaire, souvent utilisé pour le contrôle qualité d'échantillons d'ADN ou d'ARN, la précision peut atteindre la paire de base. Cependant, il n'est pas possible de séparer des molécules possédant une séquence différente, la différence de vitesse de migration étant trop faible. Cependant, les méthodes suivantes, tout en reposant sur le principe de l'électrophorèse, permettent de s'affranchir de cette limitation (Gao and Tao, 2012). On peut séparer les techniques sans séquençage les plus courantes en deux catégories. L'ARISA (Automated Ribosomal Intergenic Spacer Analysis) et la TRFLP (Terminal Restriction Fragment Lenght Polymorphism) s'appuient sur des

différences de longueur des fragments amplifiés. La DGGE (Denaturing Gradient Gel Electrophoresis) et la SSCP (Single-Strand Conformation Polymorphism) utilisent des conditions de migration où la séquence nucléotidique influe sur la vitesse de migration, et permettent donc de séparer les fragments en fonction de leur séquence.

En SSCP, les molécules double-brin sont d'abord dénaturées pour obtenir de l'ADN simple brin. Sous l'effet des liaisons faibles intramoléculaires, l'ADN simple brin adopte une conformation 3D, très dépendante de la séquence des nucléotides : une modification d'une seule paire de base peut amener à une conformation complètement changée, aux propriétés de migration différentes. Deux molécules de même taille mais de séquence différente ont une conformation simple brin différente, et peuvent alors être séparées par électrophorèse (Schwieger and Tebbe, 1998). En principe, sur les profils de migration SSCP obtenus, chaque pic correspond à une séquence. Cependant, dans la réalité, résolution de cette technique est faible : sur les profils SSCP de communautés microbiennes complexes il est fréquent d'observer de pics avec une base très large, correspondant au cumule de nombreuses séquences différentes (Haegeman et al., 2013). En utilisant un standard interne, on peut aligner les profils de différents échantillons et ainsi comparer leur diversité (analyse du profil global des courbes). Des packages d'analyse dédiés à la SSCP existent (Michelland et al., 2009b), et permettent de calculer les indices de diversité ou de construire des arbres de distance entre échantillons à partir des profils de migration.

La DGGE permet de désolidariser les deux brins d'une molécule d'ADN à un point du gel dénaturant (porteur d'un gradient d'agent dénaturant) qui dépend de la composition en bases de la séquence, les liaisons GC étant plus fortes que les liaisons AT (Muyzer, 1999). La molécule dénaturée ayant des propriétés de migration beaucoup plus faibles, deux fragments de même longueur migrent ensembles jusqu'à ce que celui comportant le moins de GC se dénature et soit freiné. Un des avantages de cette technique, qui a beaucoup été utilisée, est la possibilité de récupérer les bandes du gel et de les faire séquencer individuellement pour identifier les espèces séparées. Cependant, elle est parfois peu résolutive, certains fragments différents pouvant être mal séparés.

D'autres techniques sans séquençage existent , telles que la TRFLP ou la DHPLC (Denaturing High Performance Liquid Chromatography) mais avec le développement des nouvelles technologies de séquençage, elles ne sont maintenant que très peu utilisées pour des études de diversité microbienne (Nocker et al., 2007). Toutes ces techniques ont certains

avantages le séquençage : les temps de préparation, d'analyse puis d'interprétation sont très courts, et elles peuvent être mises en œuvre à bas coût même avec un nombre très faible d'échantillon à traiter en parallèle. Elles restent donc des outils de suivi ou de diagnostic parfois intéressants quand le délai d'obtention des résultats est un point critique. Elles sont par contre une faible résolution et n'informent pas sur la composition précise de la communauté microbienne.

III.2.2 Techniques de séquençage

.III.2.2.1 Séquençage Sanger

Historiquement, la méthode développée par Frederick Sanger a constitué la première génération de techniques de séquençage. Elle repose sur l'utilisation de nucléotides modifiés qui quand ils sont incorporés, empêchent la poursuite de l'élongation. Initialement les nucléotides étaient marqués radioactivement, mais on utilise maintenant des marquages fluorescents. Une seule molécule peut être séquencée à la fois. L'échantillon d'ADN purifié est divisé en quatre fractions, auxquelles est ajouté un seul des quatre nucléotides marqués, et qui vont toutes subir une transcription par une ADN polymérase. Celle-ci s'arrête quand elle incorpore un nucléotide modifié, et après électrophorèse on obtient les tailles de tous les fragments se terminant par un nucléotide donné. En combinant les résultats obtenus pour les quatre types de nucléotide, la séquence est reconstituée.

Les progrès en automatisation et en électrophorèse ont accompagné le développement du séquençage Sanger, qui a été l'outil principal du séquençage du premier génome humain *(Collins et al., 2003)*. Même s'il est assez lourd à mettre en œuvre, il reste à ce jour la technique la plus reconnue pour le séquençage de longs fragments purifiés, et est la technique de référence pour valider les résultats obtenus par les nouvelles technologies.

.III.2.2.2 Le séquençage nouvelle génération (Next Genation Sequencing)

Les technologies NGS diffèrent fondamentalement du séquençage Sanger sur un point : le séquençage en parallèle de séquences différentes, provenant soit d'un échantillon non purifié, soit du mélange de différents échantillons. Cette approche permet de réduire les coûts et d'augmenter le débit, mais également de séquencer des échantillons

environnementaux, ce qui a permis le développement de l'écologie moléculaire à plus grande échelle. Les NGS sont dominés par deux grandes technologies principales, aux performances, débits, précision et coûts comparables.

.III.2.2.3 Roche 454 system

La technologie 454 est la première à avoir été commercialisée. Elle repose sur la détection des pyrophosphates libérés lors de l'incorporation des nucléotides pendant la réaction de séquençage par synthèse, d'où le nom de pyroséquençage. Après une étape de PCR en émulsion, les nucléotides sont ajoutés un type à la fois au milieu réactionnel contenant les autres réactifs nécessaires à l'élongation et à l'émission d'un signal lumineux par la luciférase en réponse à la libération de pyrophosphate. L'intensité du signal est fonction du nombre de bases successives qui correspondent au nucléotide ajouté. Les nucléotides non consommés sont dégradés par une apyrase, le type suivant de nucléotide est ajouté et le signal correspondant est capturé (Liu et al., 2012a).

Les premiers séquenceurs 454 étaient capables de produire 200 000 lectures de 100 à 150 paires de base (pb). Le développement du 454 GS FLX Titanium a grandement augmenté la taille des lectures qui peuvent atteindre 700 pb, avec une précision estimée à 99,9% (une erreur toutes les 1000 bases). Cette taille des lectures est l'une des forces du 454, que pendant longtemps a été la seule technologie à proposer des lectures de cette taille. La durée du séquençage (une dizaine d'heures) était également un de ses avantages. De plus, des nombreux outils bioinformatiques dédiés à l'écologie moléculaire ont été développés pour traiter des données 454. Cependant, la technologie 454 est actuellement sur le déclin. Son débit est maintenant faible face à d'autres technologies qui ont évolué et gagné en rapidité, précision et longueur des lectures. Également, le coût par base a fortement diminué avec les nouvelles technologies de séquençage (Liu et al., 2012a; Quince et al., 2011).

.III.2.2.4 Illumina GA/HiSeq System

Le séquençage Illumina (initialement dénommé Genome Analyzer de Solexa) utilise une technologie de séquençage par synthèse. L'ADN est dénaturé en molécules simples brins qui sont alors fixées à un support (la flowcell) où est réalisée une amplification par ponts

avant que la réaction de séquençage par synthèse ne commence (Figure 108). Une étape de dénaturation permet d'attacher des fragments simples brins à la flowcell. L'extrémité libre peut alors former un pont en se liant à une amorce complémentaire, et l'ajout de nucléotides et d'enzymes permet de démarrer l'amplification par ponts (35 cycles), à l'origine de la formation des 'clusters' (ou paquets) de séquençage, ensembles de molécules identiques spatialement localisées au même endroit. Le séquençage par synthèse peut alors commencer : un nucléotide est ajouté à chaque cycle et les bases non consommées sont lavées. Une excitation laser permet d'identifier la base ajoutée à chaque cluster, et le cycle recommence.

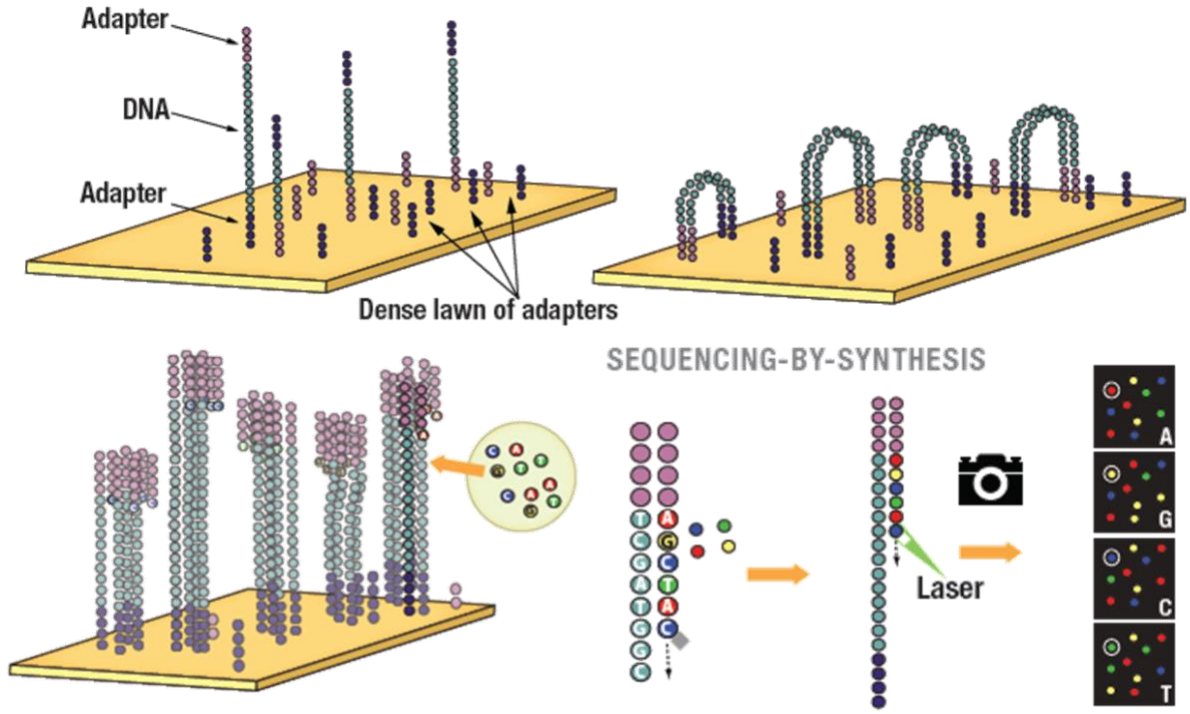


Figure 10 : Amplification par ponts et séquençage par synthèse Illumina (d'après le DOE-JGI, ressource web)

Contrairement au 454, les nucléotides sont modifiés avec un groupement fluorescent différent pour chaque base, et un groupement amovible bloquant l'élongation. Les quatre bases sont donc ajoutées simultanément, mais une seule à la fois, ce qui évite le problème des homopolymères. Le signal fluorescent est détecté en réponse à une excitation par laser, les groupements fluorescents et bloquants l'élongation sont retirés et le cycle suivant commence. Un score de qualité est calculé à partir du signal émis qui estime la certitude avec laquelle une base a été incorporée. La densité de clusters a un impact négatif sur la qualité du séquençage:

si les clusters sont trop proches les uns des autres, leurs signaux respectifs sont perturbés par ceux des voisins, induisant une diminution de la qualité. La densité de clusters est donc un paramètre critique du séquençage, et elle est déterminée par la concentration de la librairie déposée sur la flowcell.

À l'origine, les séquenceurs Genome Analyzer avaient une capacité de 1 Gigabase par run, qui a progressivement augmenté jusqu'à atteindre 85 Gigabases pour run. Lors du lancement de l'HiSeq, qui équipe aujourd'hui la plupart des plateformes de séquençage, un run de séquençage pouvait produire 600 Gigabases en une semaine. La toute dernière version de l'HiSeq en produit 1 Terabases mais surtout l'évolution des réactifs a permis d'augmenter la longueur des lectures. Par ailleurs, même si celles-ci restent courtes par rapport au 454, des techniques de lectures pairées (paired-end reads) permettent d'obtenir des lectures en tandem, chevauchantes ou non selon les protocoles. Les lectures non chevauchantes permettent d'augmenter virtuellement la longueur de lecture et sont extrêmement utiles en assemblage des génomes (Muggli et al., 2015). En séquençage de marqueurs moléculaires, l'utilisation de lectures complètement chevauchantes permet d'avoir des données de meilleure qualité, alors que l'utilisation de lectures partiellement chevauchantes permet d'obtenir des lectures plus longues (Werner et al., 2012).

Des progrès ont également été effectués pour outrepasser les effets négatifs de la richesse en GC sur les amplifications ou le séquençage (Naz and Fatima, 2013), même si certains génomes ou parties de génomes ne peuvent toujours pas être séquencés. Des séquenceurs 'de paillasse' ont également vu le jour ; ils ont des capacités plus faibles mais suffisantes pour le séquençage de marqueurs ou de petits génomes et ne nécessitant pas de grosses infrastructures. C'est le cas du MiSeq Illumina, qui est un petit séquenceur capable de produire jusqu'à 20M de séquences en 2x300pb avec les dernières versions de réactifs. En écologie moléculaire, une telle profondeur est une amélioration significative comparée à la capacité du 454. De plus, le séquençage simultané d'échantillons différents est facilité par les possibilités de multiplexage proposées par les kits Illumina. Au moment de l'amplification du marqueur par PCR, le multiplexage classique consiste à ajouter de part et d'autre de la région à séquencer un index (un code de quelques nucléotides) unique pour chaque échantillon. Après séquençage, le début et la fin de chaque séquence permettent donc de connaître l'échantillon d'origine. Cette approche nécessite cependant d'avoir autant d'amorces PCR que d'échantillons, puisque chaque amorce doit comporter la région complémentaire de la cible et un index unique. Elle a également l'inconvénient de réduire la taille du fragment séquencé,

puisque celui-ci inclut l'index. Le protocole de multiplexage Illumina propose un séquençage en trois phase de lectures (Figure 119) et permet d'utiliser un seul couple d'amorce, dont une partie est complémentaire de l'adaptateur de séquençage (nécessaire à la fixation sur la flowcell) et porte l'index, ajoutés lors d'une deuxième amplification. La région ciblée (ici ligaturée à des adaptateurs) est amplifiée par deux amorces universelles additionnées de l'adaptateur P5, et d'une séquence complémentaire de l'amorce de la 2e PCR, contenant l'index et l'adaptateur P7. La construction finale porte ainsi les adaptateurs de séquençage P5 et P7 à ses extrémités, et un index, unique par échantillon. Lors du séquençage, l'adaptateur P7 est fixé à la flowcell, et le séquençage par synthèse commence. Une fois la lecture sens 1 et la lecture de l'index terminées, c'est l'adaptateur P5 qui est fixé à la flowcell et la lecture sens 2 peut avoir lieu. Cette approche permet également de s'affranchir de biais d'amplification liés à l'index qui ont pu être constatés (Berry et al., 2011).

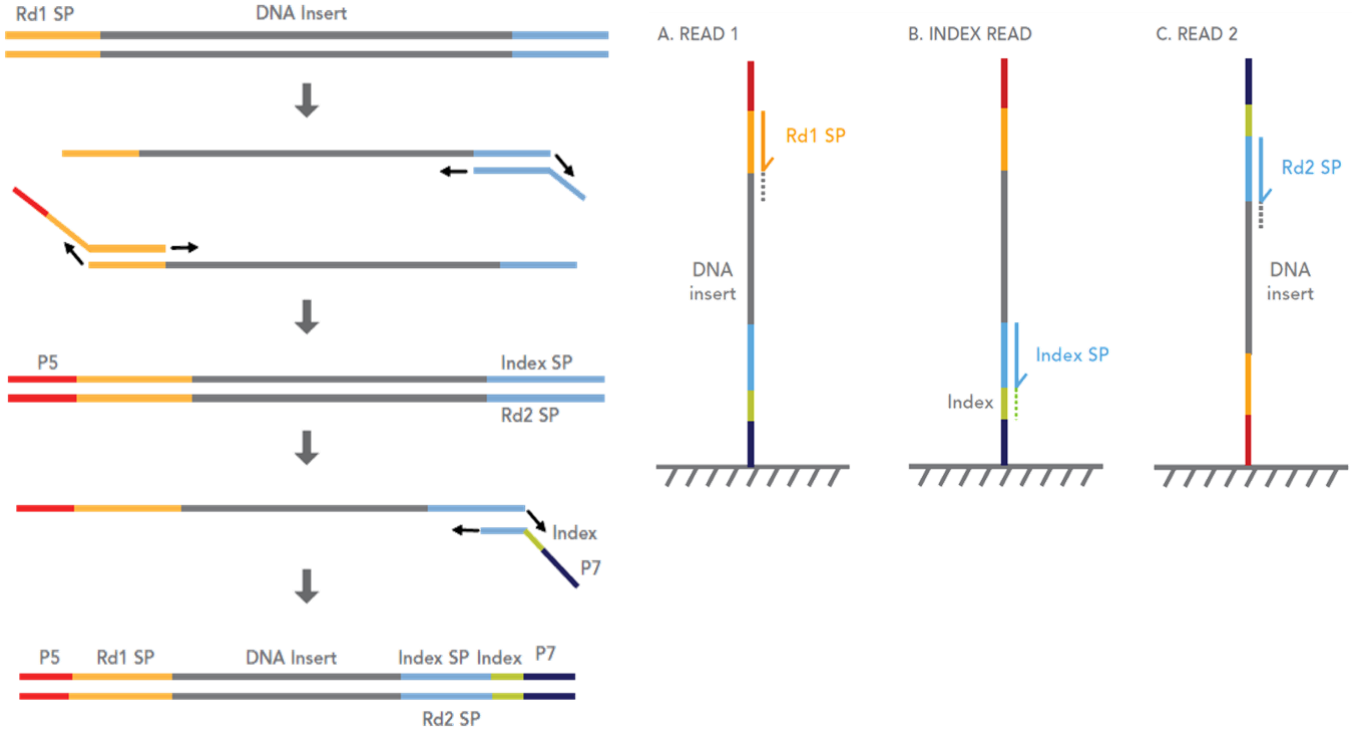


Figure 11 : Construction et séquençage d'une librairie multiplex Illumina (d'après la documentation Illumina).

Avec cette technologie, le taux d'erreurs de substitution sont assez faibles, allant de 0,006 pour la lecture sens 1 à 0,01 pour la lecture sens 2, avec une légère augmentation le long de la lecture (Schirmer et al., 2015). Cette erreur peut provenir de la réaction de séquençage par synthèse mais peut également avoir été générée lors de toutes les phases

d'amplification qui ont précédé, pendant lesquelles les polymérases sont susceptibles d'introduire des erreurs (Chen, 2014). Ils peuvent donc être différents selon les plateformes et les réactifs utilisés pendant la préparation de la librairie. Les taux d'erreurs d'insertion ou délétion sont extrêmement bas, inférieurs à 10^{-4}., ce qui évite le retraitement de données occasionné par les décalages induits pas ces délétions-insertions. Ceci est un avantage supplémentaire de l'Illumina sur le 454, dont les erreurs d'homopolymères sont des erreurs d'insertions-délétions.

La technologie Illumina a tout de même ses défauts. La qualité des séquences chute en général fortement en fin de lecture, ce qui peut être problématique puisqu'en séquençage paired-end, ce sont les fins de lectures qui sont chevauchantes et servent à joindre les deux lectures. Si les séquences en fin de lecture comportent des erreurs, la complémentarité n'est alors plus suffisante pour recréer la séquence complète. Enfin, dans le cas de séquençage multiplex, de plus en plus de travaux signalent la présence de séquences non attendues dans des échantillons parfaitement connus (Nelson et al., 2014). Deux sources ont été trouvées à ces contaminants. La première est la persistance dans le séquenceur de molécules provenant du run précédent. Ce défaut a depuis été corrigée par Illumina en modifiant les protocoles de maintenance du séquenceur entre deux utilisations. La deuxième source de contamination est la mauvaise attribution des séquences aux échantillons d'origine du, entre autres, à la forte intensité du signal de certaines séquences provenant d'un autre échantillon inclus dans la même librairie. Dans l'étude de Nelson et al. (2014), les librairies ont été préparées indépendamment et une contamination expérimentale est donc exclue ; la seule explication possible est l'existence d'erreurs de séquençage (de réaction ou d'analyse d'image) lors de la lecture de l'index. L'abondance de ces séquences 'contaminantes' atteint 0,06% (soit 12K lectures dans un run de 20M) mais dépend des runs et sans doute de leur densité en clusters. L'effet de cette source d'erreur est faible pour les analyses qui tiennent compte de l'abondance, mais contribue à la surestimation de la diversité et peut avoir un effet très fort sur des analyses basés sur la présence-absence.

.III.2.2.5 Autres technologies

Le Ion Personal Genome Machine de Ion Torrent est un nouveau séquenceur 'de paillasse' qui détecte la variation de pH associée à la libération d'un proton lors de l'incorporation d'un nucléotide par la polymérase. Les nucléotides étant ajoutés type par type,

le séquenceur détecte s'il y a addition de nucléotide et si oui, de combien, et passe à la base suivante, de la même manière qu'en 454. Cette technologie basée sur la variation du pH est peu coûteuse et extrêmement rapide (2h de run pour des fragments de 200pb) comparée aux technologies basées sur la fluorescence. Elle peut convenir parfaitement pour du diagnostic, du séquençage de plasmides ou petits génomes et éventuellement pour de petits marqueurs moléculaires.

Le séquençage Nanopore est une technologie de 3e génération, qui ne nécessite pas d'amplification par PCR et repose sur une capture en temps réel du signal associé au séquençage. De nombreux puits sont percés formant un canal unique d'heptamères d'α-haemolysin. Un courant ionique continu permet d'établir une différence de potentiel entre les deux côtés de la membrane percée par le canal protéique. Toute perturbation de ce canal, notamment le passage à travers lui d'une molécule d'ADN simple brin, affecte la différence de potentiel qui peut être mesurée par des techniques classiques d'électrophysiologie. La forme de chaque base étant différente, leur passage à travers le canal est discernable et permet donc une lecture de la molécule sans aucune synthèse et ne nécessitant pas de réactifs. Cette technologie, qui permettrait le séquençage de fragments de 5k pb à une vitesse d'une base par nanoseconde, avec un séquenceur de la taille d'une grosse clé USB, n'est cependant pas encore commercialisée.

Le SMRT PacBio (Single Molecule Real-Time Sequencing de Pacific BioScience, couramment appelé PacBio) est également un séquenceur de 3e génération. Il est également capable de séquencer des molécules uniques : chaque molécule est isolée des autres dans un puit où est présente une seule polymérase. Cette polymérase est modifiée pour couper le groupement fluorescent associé aux bases au moment de l'incorporation. Le signal associé est détecté en temps réel par une caméra, qui permet de déduire la séquence grâce à la nature du signal (Eid et al., 2009), mais aussi d'éventuelles modifications structurales de l'ADN comme sa méthylation, qui a un impact sur le délai entre l'incorporation de plusieurs bases (Clark et al., 2011). Contrairement au Nanopore, cette technologie est en service. Elle permet de séquencer jusqu'à 150 000 fragments de 8 à 12k pb en quelques heures. De tels longs fragments peuvent grandement faciliter l'assemblage, jusqu'ici très difficile, des régions répétées ou à la complexité faible dans les génomes. Enfin, en s'affranchissant des biais d'amplification des techniques de 2e génération, le séquençage PacBio pourrait également enfin résoudre le problème des régions et génomes riches en GC, jusqu'ici non séquencés (Shin et al., 2013).

III.3. Analyse des données amplicon générés par NGS

En sortie de séquenceur, les données brutes obtenues prennent la forme de fichier FASTA, succession de lignes d'en-tête et de séquences ADN, ou FASTQ, composés d'un en-tête, de la séquence ADN et du score de qualité associé à chaque base. Ces données nécessitent des étapes de traitement pour obtenir des tables d'abondance qui soient les plus proches de la réalité des échantillons séquencés. Ce traitement doit permettre de convertir les données brutes en données biologiques et doit, autant que possible, prendre en compte les biais générés par les méthodes de séquençage afin de les corriger.

III.3.1 Étapes clés et outils dédiés

Le traitement de données de séquençage de marqueurs moléculaires peut se diviser en quatre grandes étapes, communes aux différentes technologies de séquençage. De nombreux outils différents y sont souvent associés, et il est parfois difficile de savoir comment décider quel outil utiliser quand plusieurs sont disponibles pour une même tâche. Les outils les plus communs et/ou les plus performants seront présentés ici de manière non exhaustive.

.III.3.1.1 Nettoyage des données

La première étape du traitement des données consiste à ne récupérer que l'information fiable et de qualité. Dans le cas du séquençage multiplex, il est nécessaire de retirer les index ajoutés à la région séquencée, ainsi que les régions complémentaires des amorces de PCR. En effet, pendant l'amplification, quand la région cible n'est pas strictement complémentaire de l'amorce, l'hybridation a tout de même lieu et au cours des cycles PCR suivants, la séquence réelle disparait au profit de la séquence de l'amorce. En fin de PCR, toutes les molécules amplifiées sont ainsi porteuses de la séquence de l'amorce qui ne corresponde pas à la séquence d'origine. Il est donc nécessaire d'éliminer la partie des séquences correspondant aux amorces. Cette étape est assez simple à réaliser à l'aide d'outils dédiés. Le plus commun pour cet usage est Cutadapt (Martin, 2011), initialement développé pour retirer les séquences d'adaptateur qui peuvent se retrouver dans les données brutes quand on séquence de très petits fragments. Cutadapt est capable de gérer les fichiers Fasta et Fastq et tolère les bases dégénérées dans les amorces (ou adaptateurs).

C'est également pendant l'étape de nettoyage que les données peuvent être filtrées selon différents critères de qualité. Il est possible avec des fichiers fastq, porteurs d'un score de qualité pour chaque base, d'utiliser des filtres basés sur la moyenne de ce score pour une séquence, ou sur sa moyenne dans une fenêtre glissante. D'autres critères plus simples sont corrélés à ce score de qualité, comme la présence d'erreurs dans les zones d'amorces ou la présence de bases indéterminées (N) dans les séquences. En 454, des longueurs de séquence faibles ou des séquences comportant des homopolymères de grande taille sont également des indices de mauvaise qualité d'une séquence. Toujours en 454, les erreurs d'homopolymères peuvent être réduites en retravaillant le signal de la sortie brute par des outils comme PyroNoise (Quince et al., 2009) puis AmpliconNoise (Quince et al., 2011).

Pour des données Illumina pairées, il est également nécessaire de réassembler la paire de séquences en une séquence unique. Le bon alignement de la zone chevauchante joue alors un rôle de filtre qualité, puisque si des erreurs se sont accumulées en fin de lectures, celle-ci comporte alors beaucoup de mauvais appariement, dont la fréquence est un critère essentiel pour les assembleurs comme Flash (Magoč and Salzberg, 2011) ou PANDAseq (Masella et al., 2012). Généralement, avec un bon paramétrage de ces outils, un filtre qualité ultérieur n'a pas ou peu d'effet sur le jeu de données.

.III.3.1.2 Déchimérisation

Les chimères sont des séquences artéfactuelles qui ont pour origine le décrochement d'une polymérase au cours d'un des cycles d'élongation et l'hybridation de la séquence incomplète sur un brin matrice différent lors d'un des cycles suivants. Il en résulte donc une séquence chimérique composée de parties de deux séquences réelles ou plus (Figure 12). Les chimères sont des sous-produits de PCR et sont donc relativement rares (de 5 à 10% selon les estimations), mais ont un impact extrêmement fort sur la diversité et la richesse observée. Les chimères n'ont aucune existence réelle dans l'échantillon de départ et doivent donc être détectées et éliminées. D'après certains auteurs, elles pourraient être à l'origine de la biosphère rare, qui serait donc construite d'artéfacts (Kunin et al., 2010).

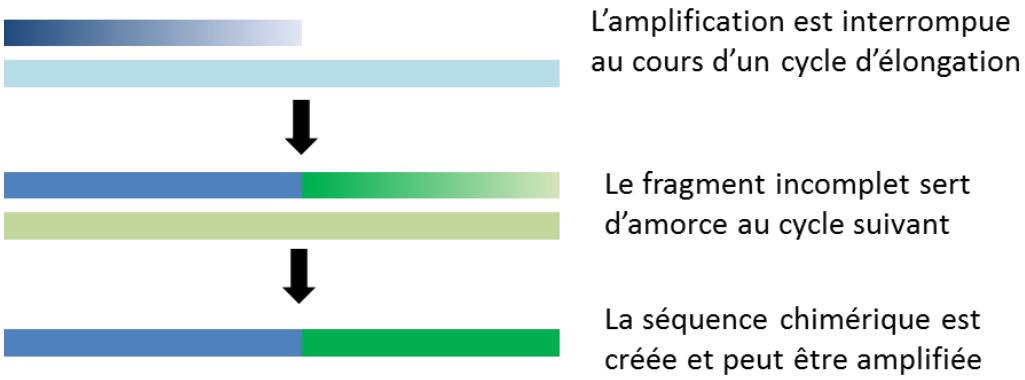


Figure 12 : Principe de formation des chimères de PCR.

Plusieurs outils permettent de détecter ces chimères, ou en tout cas une partie d'entre elles. Ils reposent sur deux approches différentes, l'une s'appuyant sur une base de données qui sert de référence, et l'autre qui utilise l'échantillon lui-même comme auto-référence. Les outils les plus connus, ChimeraSlayer (Haas et al., 2011) et UCHIME (Edgar et al., 2011) sont capables de travailler selon les deux méthodes.

L'approche « base dépendante » consiste à découper la séquence à tester en plusieurs fragments, qui seront soumis indépendamment à une recherche de meilleur alignement contre la base de donnée. Si les différents fragments s'alignent avec de très bons scores sur des séquences différentes de la référence, la séquence testée est chimérique. Les outils diffèrent par leur façon de rechercher les séquences avec la plus forte similarité (par exemple BLAST pour ChimeraSlayer), leur méthode d'alignement et enfin, leur calcul du score permettant de déterminer si une séquence est ou non chimérique.

L'approche « base indépendante », ou « *de novo*, » repose sur les mêmes principes, à la différence que la base de données est construite à partir des séquences de l'échantillon. Les chimères étant des sous-produits d'amplification, elles sont nécessairement en plus faible abondance que les séquences mères qui ont servi à leur création. Les séquences les plus abondantes d'un échantillon ne sont donc à priori pas chimériques. C'est sur cette caractéristique que repose la détection *de novo* : les deux séquences les plus abondantes de l'échantillon sont considérées comme non chimériques, et constituent la banque de référence interne de départ. La 3^{e} séquence est alors testée contre cette banque interne. Si elle n'est pas détectée comme chimérique, elle est ajoutée à la banque. Toutes les séquences sont ensuite testées itérativement de la même manière. Le calcul de score est légèrement modifié et prend

également en compte le rapport entre abondance de la séquence testée et abondance de ses parents potentiels, qui doit dépasser une valeur seuil.

La détection de chimère n'est pas infaillible et peut ne pas détecter une séquence alors qu'elle est chimérique. Un paramétrage plus agressif des outils, et notamment du calcul de score, peut permettre d'augmenter la sensibilité et donc le nombre de chimères détectées, mais il augmente aussi le nombre de faux positifs (séquences détectées comme chimériques alors qu'elles ne le sont pas). Il est donc nécessaire de trouver un compromis entre la détection d'un maximum de chimères et d'un minimum de faux positifs.

Alors que UCHIME reste l'outil le plus utilisé, ses auteurs le considèrent obsolète en faveur de UPARSE (Edgar, 2013), un algorithme plus efficace qui réalise la détection de chimère en même temps que le regroupement (clustering) des séquences en unités taxonomiques opérationnelles ou OTUs (operational taxonomic units). Cependant, par cette méthode la détection de chimère ne peut pas être utilisée indépendamment du clustering et ne peut donc pas être utilisée avec d'autres outils que UPARSE.

La détection des chimères reste un point critique des pipelines de traitement de données. Des outils continuent d'être publiés et comparent leurs résultats. CATCh (Combining Algorithms to Track Chimeras) utilise cinq des outils les plus reconnus, et combine leurs résultats pour augmenter la sensibilité et précision de sa détection (Mysara et al., 2015). Cependant, son exécution est assez difficile et les temps de traitements sont longs du fait de l'utilisation de divers algorithmes parfois assez lents.

.III.3.1.3 Clustering

Deux approches peuvent être envisagées pour obtenir des tables d'abondances classifiées : l'assignation taxonomique directe de toutes les séquences, suivie du regroupement des séquences ayant la même assignation, ou le regroupement en 'paquets' (clustering) non supervisé des séquences, suivi d'une assignation taxonomique des clusters formés. La première approche est la plus simple à mettre en œuvre, elle est entièrement parallélisable et les temps de calculs n'augmentent que linéairement avec la taille des données. Cependant, elle présente des défauts majeurs. Les séquences absentes de la référence ne sont pas assignées et elles sont regroupées dans une catégorie « non classifié » qui peut alors contenir des séquences très différentes. De plus, la méthode ne fournit aucune information sur sa structure interne (combien de bactéries différentes, proches entre elles ou

non…). Cette approche peut être suffisante dans le cas du séquençage d'échantillons très bien connus, mais elle n'est pas appropriée pour l'analyse de séquences nouvelles ou inconnues. De plus, elle peut regrouper des taxons différents si l'outil d'assignation taxonomique n'est pas assez précis : deux espèces d'un même genre seront regroupées en une seule catégorie si l'outil d'assignation n'est pas capable de déterminer l'espèce. La richesse en taxons ne peut donc pas être évaluée par cette méthode, alors qu'elle est prise en compte dans le calcul de la plupart des indices de diversité. Pour toutes ces raisons mais également pour prendre en compte le bruit de séquençage et la variabilité intragénomique et intraspécifique, les méthodes avec clustering sont classiquement préférées.

Dans les approches de clustering, les séquences sont regroupées en OTUs à l'aide d'un seuil de distance entre séquences. Différents seuils ont été établis à partir de l'analyse de séquences d'ARNr complètes (Hugenholtz et al., 1998). Ils sont reconnus pour assez bien correspondre aux espèces (1-3% de différence), genres (<5% de différence) ou classes (<15%). Le clustering permet alors la détection d'unités taxonomiques dans des environnements inconnus et complexes, même si une assignation taxonomique précise ne peut leur être donnée.

À l'origine, les approches de clustering débutent par la construction d'un alignement multiple des séquences d'ARNr 16S. Cet alignement est alors utilisé pour calculer les distances deux à deux entre toutes les séquences, ces distances étant exprimées en pourcentage de nucléotides différents entre les deux séquences. La matrice de distance ainsi formée est alors utilisée par un algorithme de clustering hiérarchique. Une séquence est choisie pour être le « centre » du premier OTU (les méthodes de choix et la définition de centre varient selon les outils) et toutes les séquences qui en sont assez proches (en fonction du seuil choisi) sont ajoutées au premier OTU. Une autre séquence est alors choisie et sert de centre à l'OTU suivant. Cependant, les algorithmes de clustering hiérarchique classiques (DNAClust (Ghodsi et al., 2011), CD-Hit (Fu et al., 2012), ou Uclust (Edgar, 2010)) présentent deux principaux défauts, l'utilisation d'un seuil global et une dépendance à l'ordre des séquences. Les résultats produits sont évidemment très dépendants du seuil fixé, alors que celui-ci, s'il n'est pas totalement arbitraire, est une approximation qui a ses limites. En effet, les espèces bactériennes n'évoluent pas toutes à la même vitesse, et aucun seuil n'est vrai pour toutes les branches de l'arbre phylogénétique des bactéries. Un seuil global est donc inévitablement trop large pour des taxons à l'évolution lente (Nebel et al., 2011), donc aux séquences peu variables, et trop restreint pour les taxons à évolution rapide. La conséquence

de l'utilisation d'un seuil inadapté est la fusion de taxons proches à la séquence peu variable, et la scission en nombreux OTUs de taxons à forte variabilité. L'autre faiblesse du clustering hiérarchique est une influence de l'ordre d'analyse des séquences par l'algorithme : une séquence peut être exclue d'un OTU alors qu'elle y aurait été incluse si le clustering était parti d'un autre centre (Figure 13).

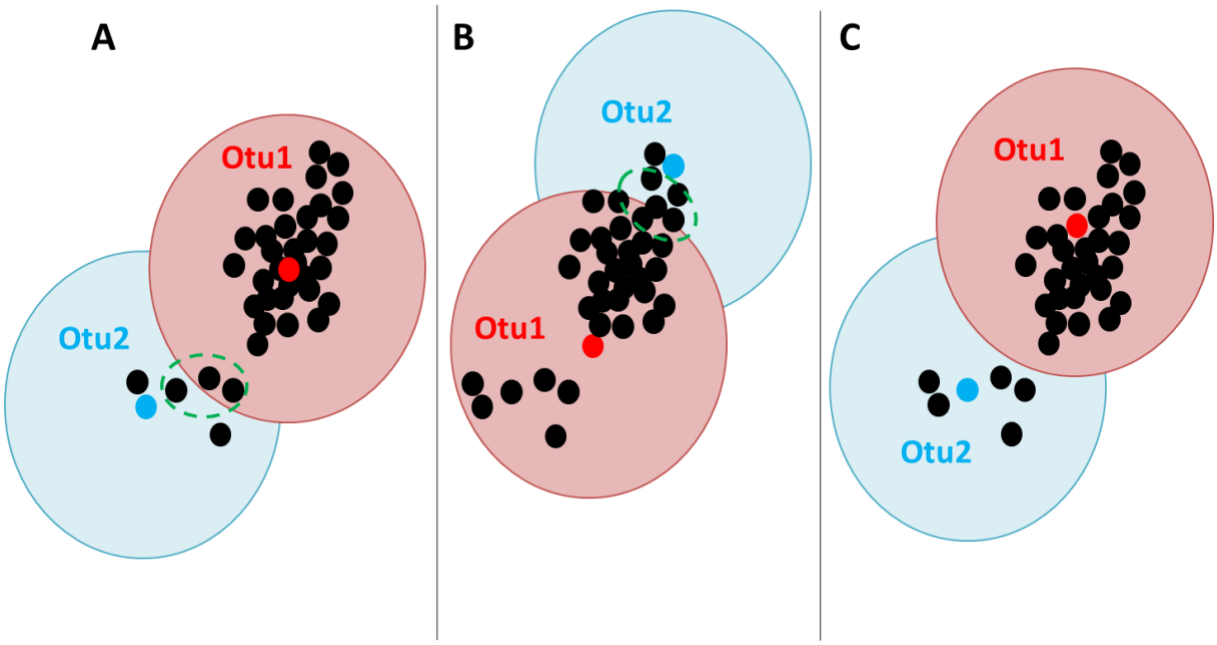


Figure 13 : schématisation de la formation des OTUs par un outil de clustering hiérarchique classique. Les séquences sont représentées par les cercles noirs, la distance entre eux traduisant la distance entre séquences. Les séquences en rouge et bleu correspondent au centre choisi pour l'OTU1 (cercle rouge) et l'OTU2 (cercle bleu) dans chaque cas A, B et C. En fonction des centres choisis, les OTUs construits sont différents. Dans le cas A et le cas B, des séquences très proches entre elles sont séparées dans des OTUs différents (cercles pointillés verts).

Les séquences attribuées à un OTU n'étant plus considérées lors de la création des OTUs suivants, le clustering hiérarchique peut amener à séparer de séquences pourtant très proches entre elles (Figure 13, cercles verts), ce qui amène à des cas de mauvaise construction d'OTUs qui ne sont ni rares, ni restreints à quelques séquences (Koeppel and Wu, 2013).

Avec l'augmentation de la taille des jeux de données due aux progrès des technologies de séquençage, ces algorithmes, en plus de leurs défauts, atteignent parfois leurs limites techniques du fait d'une croissance exponentielle du temps de calcul et des ressources en mémoire requises. Quelques algorithmes comme Esprit-Tree (Cai and Sun, 2011) permettent

une pseudo-parallélisation du clustering et ont permis de repousser les limites du clustering hiérarchique en subdivisant le jeu de données, mais ne sont pas capables de supporter l'ensemble des donnés produits en un run MiSeq. Avec l'augmentation de la profondeur de séquençage, on observe une augmentation continue de la richesse observée, qualifiée de biosphère rare (Pedrós-Alió, 2012). Celle-ci est toujours l'objet de débats entre les auteurs qui la considèrent réelle voir même sous-estimée (Gonzalez et al., 2012), et ceux qui la considèrent artéfactuelle (Huse et al., 2010) puisqu'on la détecte même dans le cas d'échantillons connus à la diversité contrôlée (mélanges connus de souche ou mock communities).

UPARSE (Edgar, 2013), publié en 2013, est le premier outil à fournir des résultats proches des résultats attendus sur d'échantillons de composition connue. En effet, avec tous les autres outils existants, la diversité était systématiquement surestimée. UPARSE propose une approche qui en associant la mise à l'écart des lectures singleton (présentes une seule fois dans l'ensemble du jeu de données) et un algorithme qui réalise simultanément clustering et détection de chimères, permet d'obtenir un nombre d'OTU très proche de l'attendu là où les autres méthodes en donnent 5 à 10 fois plus. Le filtrage des lectures singleton est optionnel mais proposé par défaut, il permet une forte réduction du nombre d'OTUs détectés mais participe également beaucoup à l'exceptionnelle rapidité d' UPARSE. En effet, en plus de sa précision validée sur des mélanges contrôlés de souches, UPARSE est capable de traiter en quelques heures des jeux de données de grande taille, dont le traitement pouvait prendre plusieurs semaines avec d'autres outils. Cependant, il ne résout pas les problèmes liés au clustering hiérarchique à seuil global présentés précédemment (Figure 13).

Swarm utilise une approche complètement différente, qui s'affranchit du choix d'un seuil global et n'est pas dépendante de l'ordre d'analyse des séquences (Mahé et al., 2014). Un OTU y est défini comme un ensemble de séquences reliées entre elle par des 'chaînes' de séquences ne présentant qu'une seule différence (généralisable à *d* différences ; Figure 14). Autrement dit, un OTU est un ensemble continu de séquences séparées par *d* différences. Cette approche repose sur l'hypothèse que les OTUs sont séparés par plus de *d* différences, et que les séquences intermédiaires permettant de créer un pont entre deux OTUs n'existent pas. En pratique, les OTUs sont toujours constitués d'une séquence très abondante (la graine), la séquence réelle, et d'un nuage de séquences faiblement abondantes présentant quelques différences avec la graine, correspondant au bruit de séquençage, c'est-à-dire à des erreurs.

Plus une séquence est abondante (ou plus la profondeur de séquençage augmente), plus son nuage d'erreurs est large.

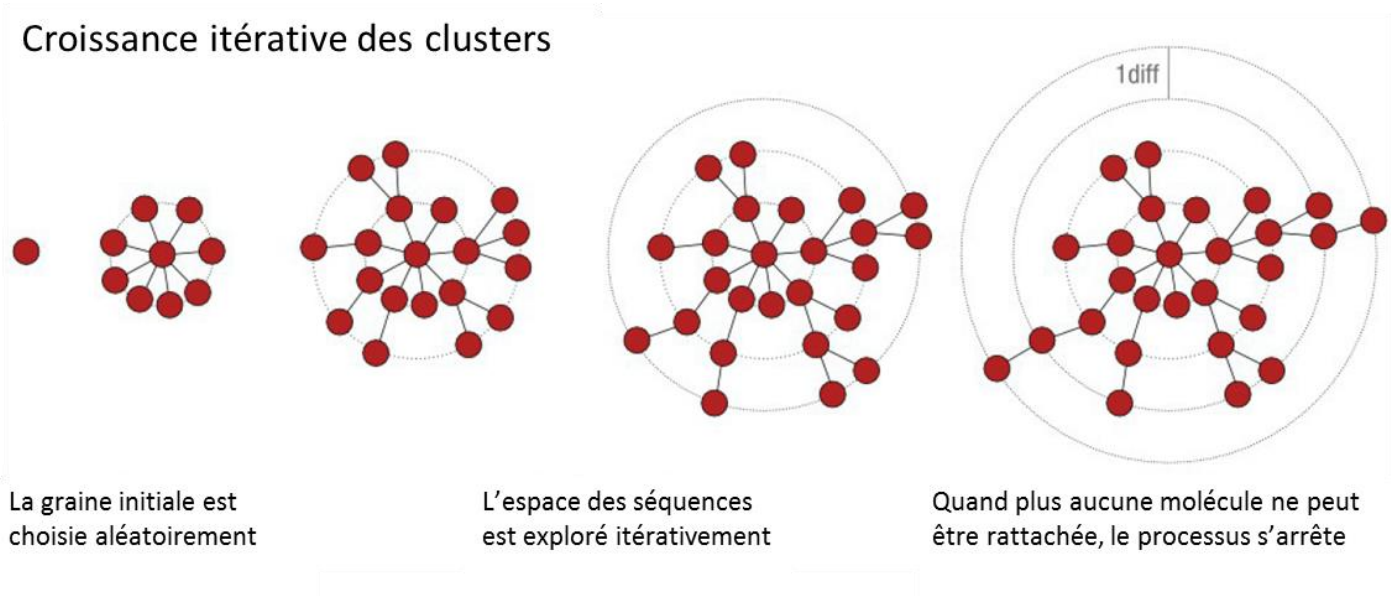


Figure 14 : Schématisation du principe de clustering de Swarm (d'après Mahe, 2014). Une séquence (cercle rouge) est choisie aléatoirement, et on recherche toutes les séquences présentant une seule différence (tiret noir) avec elle, qui sont ajoutées à la couronne c1. L'opération est alors recommencée pour trouver toutes les séquences présentant une seule différence avec n'importe laquelle des séquences de c1. Le processus est répété itérativement jusqu'à ce que plus aucune séquence ne puisse être ajoutée. L'ordre n'a ici pas d'effet sur la construction des OTUs : en démarrant depuis une extrémité de chaîne, on aboutit au même assemblage. Aucun seuil global n'intervient, Swarm permet de dessiner les limites naturelles des OTUs en fonction de la richesse de leur couronne.

Swarm n'utilise pas de seuil global et trouve les limites naturelles de chaque cluster. De plus, l'ordre des séquences n'a aucune influence sur la construction des OTUs, excepté des gains de temps de calcul en triant les séquences par abondance. Enfin, pour corriger les cas où deux OTUs très proches seraient reliés par des nuages d'erreurs chevauchants, les clusters produits sont ré-analysés à la recherche de motifs caractéristiques (deux séquences très abondantes reliées par un pont de séquences rares) et découpés le cas échéant. Le clustering Swarm est extrêmement précis et produit un nombre élevé de clusters, dont une part non négligeable de singletons qui doivent être éliminés pour retomber sur des nombres d'OTUs produits proches des valeurs attendues.

.III.3.1.4 Assignation taxonomique

Les performances des outils d'assignation taxonomique, qui permettent d'associer une séquence à une taxonomie, sont extrêmement dépendants des bases de données utilisées (voir 0 Banques de données). De manière résumée, deux approches d'assignation sont couramment utilisées, celles utilisant une recherche par Blast et celles utilisant une recherche par k-mer telle que proposé par le classeur de Ribosomal Database Project (RDP classifier : Wang et al., 2007).

Les méthodes par Blast reposent sur la recherche par des algorithmes de type Blast (ou megablast ou Blat) dans une base de référence. La taxonomie du meilleur alignement est alors donnée à la séquence assignée, et on dispose des données de couverture, pourcentage d'identité et une valeur de significativité (e-value). Une des forces de la méthode est l'accès à la couverture, qui dans le cas de séquences alignées avec 100% d'identité mais sur une partie seulement de leur longueur, peut indiquer qu'on a affaire à une chimère non détectée. Dans le cas de séquences alignées sur toute leur longueur, le score d'identité permet alors d'estimer la fiabilité de la taxonomie attribuée. Certaines variantes de l'assignation par Blast utilisent non pas le meilleur alignement, mais les *n* premiers, et attribuent à la séquence testée la taxonomie consensus des meilleurs alignements. Ces approches atteignent rarement une assignation à l'espèce et ont donc une précision plus faible, mais des taux d'erreurs plus faibles, surtout dans le cas d'environnements peu connus.

Les méthodes k-mers de type RDP Classifier sont plus rapides parce qu'elles n'ont notamment pas recours à un alignement. Pour simplifier, les méthodes k-mers reposent sur la construction d'une table de probabilité qu'un k-mer (mot de *k* lettres) soit représenté dans un taxon de la base de référence. Une fois cette table construite, une séquence est testée en y tirant au hasard des k-mers. En croisant la composition en k-mers et la table de probabilité, on obtient une taxonomie à différents niveaux (du Phylum au genre ou à l'espèce) pour la séquence. Une estimation de la fiabilité de l'assignation est fournie par un 'bootstrap' qui consiste à retirer plusieurs fois (en général une centaine de fois) au hasard des k-mers pour s'assurer que le résultat soit le même quel que soit le tirage aléatoire. La méthode a prouvé sa rapidité et son efficacité ; elle a aussi l'avantage d'être capable d'assigner des séquences à des niveaux taxonomiques très larges (phylum, classe ou ordre) quand une séquence est inconnue dans la banque de référence. Par contre, elle est quasiment aveugle aux chimères déséquilibrées, qu'elle assigne avec une certitude élevée à la séquence parent majoritaire.

Chaque approche d'assignation taxonomique a ses avantages et inconvénients, l'une comme l'autre sont couramment utilisées et acceptées et c'est surtout la banque de données utilisée qui est déterminante.

III.3.2 Outils d'analyse de séquences généraux

Tous les outils et approches présentés précédemment peuvent s'utiliser les uns avec les autres en se pliant à leurs formats des données d'entrée et sortie. Ils nécessitent une installation individuelle avec ses prérequis, tâche fastidieuse et parfois complexe voire impossible pour les utilisateurs de serveurs de calculs. D'autres logiciels incorporent en un seul un grand nombre d'outils courants et s'occupent du formatage des données entre les différents outils. C'est le cas de deux outils particulièrement utilisés, Mothur (Schloss et al., 2009) et QIIME (Caporaso et al., 2010).

.III.3.2.1 Mothur

À l'origine conçu pour traiter des données 454, Mothur incorpore des outils spécifiques du 454 comme des débruiteurs pour corriger les erreurs d'homopolymères. Mais chaque outil étant optionnel, il est complètement adaptable à des données Illumina. Les outils qu'il contient permettent de passer de données brutes à des tables d'abondances avec assignation taxonomique ; quelques banques de données dans le format compatible Mothur sont disponibles au téléchargement. Mothur propose une large gamme de choix pour la plupart des étapes clés citées précédemment. Il propose aussi l'estimation des courbes de raréfaction, des indices de diversité, des diagrammes de Venn ou autres statistiques, notamment de test d'hypothèses. Les possibilités sont très vastes, mais une des faiblesses de Mothur réside dans sa méthode de clustering. Celle-ci repose sur la construction d'une matrice de distances entre séquences qui nécessite d'être complètement chargée en mémoire pour que l'algorithme de clustering, très lent, puisse tourner. Cette étape est la seule où il n'est proposé qu'un seul outil, pour lequel sont paramétrables le seuil global ainsi que la définition de cluster (voisin le plus proche, le plus éloigné, ou moyen). Il est possible de réduire le nombre de séquences à clusteriser à l'aide d'un outil de pre-clustering qui permet de réduire le bruit de séquençage : après un alignement multiple, les séquences sont analysées et si l'une d'entre elles, très rare, présente moins de *d* différences avec une abondante, elle est considérée

comme artéfactuelle, son abondance est ajoutée à la séquence abondante et sa séquence éliminée du jeu de données. Cependant, dans le cas de jeux de données de plusieurs millions de séquences, il devient techniquement impossible d'utiliser Mothur selon la procédure standard conseillée par les auteurs (Schloss et al., 2011) du fait de cette étape limitante de clustering. Mothur est un outil disponible en ligne de commande mais il existe une interface utilisateur graphique (GUI) disponible pour Mac, utilisable uniquement en local avec de petits jeux de données.

.III.3.2.2 QIIME

De la même façon que Mothur, QIIME (pour Quantitative Insights Into Microbial Ecology) rassemble de très nombreux outils qui sont rendus interopérables par un ensemble de scripts permettant de faire correspondre entrées et sorties. Il est facilement possible de l'installer en local pour une utilisation avec de petits jeux de données. Les formats qu'ils utilisent ne sont pas les mêmes, mais les possibilités de Mothur et QIIME sont équivalentes et il n'existe pas de réelles différences de résultats entre les deux puisqu'ils reposent sur l'utilisation des mêmes outils. La différence majeure entre les deux réside dans le choix d'outils proposés pour certaines étapes : par exemple, QIIME ne dispose pas d'un outil de pre-clustering, mais inclut plusieurs outils de clustering dont UPARSE et Swarm (voir .III.3.1.3 Clustering), qui peuvent permettre de traiter des jeux de données de grande taille. QIIME est pour le moment uniquement disponible en ligne de commande mais une interface graphique utilisateur est actuellement en développement pour QIIME2.

III.4. Outils statistiques pour l'écologie moléculaire

III.4.1 Indices (diversité α)

Pour caractériser la diversité d'un écosystème, il existe des indices de diversité avec lesquels une communauté est associée une valeur numérique. Les indices de diversité sont un des premiers outils permettant de comparer la diversité de communautés, qu'elles soient proches ou totalement différentes. Il existe de nombreux indices reposant sur des méthodes de calculs différentes. La plupart ont été définis en écologie classique mais sont également

utilisables en écologie moléculaire, puisque seule la méthode d'acquisition des données change.

Les plus simples consistent en des comptages du nombre de taxons observés (richesse spécifique). Comme celui-ci peut être très dépendant de la profondeur de séquençage, d'autres indicateurs permettent d'estimer la valeur de richesse si elle atteignait son plateau, en se basant notamment sur la fréquence des singletons. C'est le cas de l'estimateur de Chao (Chao, 1987), assez couramment utilisé en écologie microbienne. Cependant, les techniques de séquençage et d'analyse actuelles produisent du bruit consistant notamment en un très grand nombre de singletons, qui s'ils ne sont pas filtrés conduisent à des surestimations très fortes de la richesse. L'application de filtres permet de corriger en partie ce bruit de séquençage, mais en contrepartie élimine la plupart du temps tous les singletons, à la base du calcul de Chao. Richesse et Chao sont donc deux éléments à considérer avec précaution, notamment s'ils servent à comparer des résultats provenant de différentes méthodes d'analyse.

Cependant, d'autres indices de diversité, prenant en compte l'abondance relative de chaque taxon, sont moins impactés par les singletons et rares et sont également plus informatifs que la richesse seule, même si celle-ci apporte une information complémentaire. Les indices de Shannon et Simpson sont les plus courants. L'indice de Shannon se calcule comme une somme pour chaque taxon d'une fonction de son abondance relative (abondance relative multipliée par son propre log2), et permet de quantifier l'hétérogénéité d'une population. Plus sa valeur est élevée, plus la population est diversifiée et les taxons équitablement abondants. Le calcul de l'indice de Simpson est assez similaire (somme pour chaque taxon de son abondance relative au carré). La valeur obtenue est rarement utilisée directement puisqu'elle est contre-intuitive : la diversité augmente quand l'indice diminue, et est maximale pour une valeur de 1/R (avec R la richesse spécifique de l'échantillon). À la place, on utilise en général l'inverse de Simpson, qui lui augmente avec la diversité et a pour maximum R, la richesse spécifique, si tous les taxons sont équi-abondants. Ces indices sont très couramment utilisés et du fait du très faible poids accordé aux taxons rares, sont en général comparables d'une étude à l'autre, même si les techniques ou méthodes d'analyse sont différentes.

Enfin, d'autres indices prennent en compte la distance phylogénétique entre taxons présents. Ainsi, la diversité phylogénétique (PD) est estimée en mesurant dans un arbre phylogénétique la somme des longueurs de branche permettant de relier tous les taxons

présents. Un très grand nombre de taxons tous très proches a ainsi une diversité phylogénétique plus faible que quelques taxons phylogénétiquement très éloignés.

III.4.2 Distances (diversité β)

Les indices de diversité permettent de caractériser la diversité α de plusieurs échantillons, mais renseignent peu sur la ressemblance entre échantillons. En effet, deux échantillons peuvent avoir la même richesse mais être composés de taxons complètement différents, ou avoir le même indice de Shannon mais correspondre à deux structures de communautés très différentes, l'une avec une très forte richesse mais des taxons dominants, l'autre une faible richesse mieux répartie. Cependant, il existe de nombreuses méthodes de calcul de distance entre échantillons. La distance euclidienne (la distance mathématique la plus simple) peut tout à fait être appliquée à une table d'abondance, mais d'autres distances plus adaptées à ce type de données sont disponibles.

L'indice de Bray-Curtis (mathématiquement un indice et non une distance) se base sur les abondances (non relatives !) de taxons des échantillons à comparer, et nécessite donc de travailler sur des échantillons de profondeur comparable. Son calcul fait intervenir le rapport minimum/somme pour chaque taxon des échantillons comparés. Si tous les taxons sont également abondants entre échantillons, l'indice vaut 0. Si tous les taxons sont spécifiques à un seul des échantillons (abondance 0 dans un des deux échantillons), l'indice vaut 1. L'indice de Bray-Curtis est un des indices les plus utilisés pour comparer des communautés. Il est souvent considéré comme une distance et les matrices de pseudo-distance obtenues sont utilisées par des outils d'analyse des distances.

Les autres distances les plus pertinentes en écologie microbienne sont les distances Unifrac (Lozupone and Knight, 2005; Lozupone et al., 2011). Basées sur le même principe de calcul que la diversité phylogénétique, elles font intervenir celle-ci en mesurant la longueur des branches qui ne sont pas communes entre échantillons comparés (Unifrac) ou bien en faisant la différence entre les longueurs de branches pondérées par leur abondance relative (Unifrac pondéré). Ces distances permettent de prendre en compte la distance phylogénétique entre taxons, et d'ainsi éviter que deux taxons très proches mais spécifiques chacun d'un échantillon donné, ne soient à l'origine d'une distance élevé (ce que l'indice de Bray-Curtis fait). Cependant, si la distance Unifrac pondérée est très utile et puissante pour comparer des

échantillons très différents (avec éventuellement très peu de taxons communs), la complexité de son calcul peut compliquer son interprétation.

III.4.3 Méthodes de projection et visualisation

La façon dont différents échantillons se structurent et les ressemblances entre eux peuvent être visualisées par différentes méthodes graphiques. La plus courante est l'analyse en composante principale (ACP), qui permet de projeter sur des axes maximisant la variance des données la position de chaque point. La proximité graphique entre échantillons correspond alors à une ressemblance entre eux. D'autres méthodes statistiques, comme la régression des moindres carrés partiels (PLS - Partial Least Square) repose sur le même principe mais entre deux tableaux de données, tout en maximisant la corrélation entre les deux ensembles de données. Enfin, d'autres méthodes existent (analyse de redondance, analyse de correspondances canoniques, etc) dont les spécificités ne seront pas discutées ici. Toutes les méthodes présentées, dont la très répandue ACP, sont applicables à tout type de données numériques, mais ne sont cependant pas très adaptées à des matrices de distance, et ne peut permettent pas une bonne analyse basée sur l'indice de Bray-Curtis ou prenant en compte le poids de la distance phylogénétique en se basant sur les distances Unifrac. Cependant, d'autres méthodes, similaires en pratique à l'ACP, sont adaptées spécifiquement à des matrices de distance. C'est le cas de la PCoA (Principle Coordinate Analysis ou Multidimensional Scaling – MDS) ou du NMDS, sa version non métrique.

Des solutions permettant de calculer ces distances et d'utiliser ces méthodes graphiques existent dans des package en langage R dédiés à l'analyse de communautés microbiennes (PhyloSeq : McMurdie and Holmes, 2013) ou de données –omiques (mixOmics : Cao et al., 2009).

CHAPITRE II

PROCEDURES EXPERIMENTALES

CHAPITRE II : PROCEDURES EXPERIMENTALES

I. Origine des populations bactériennes

I.1. Rumen initial

Les échantillons de rumen ayant servi de source d'inoculum aux expériences d'enrichissement en réacteur proviennent du rumen de deux vaches Holstein (en dehors d'une période de lactation) de l'INRA Clermont-Ferrand. Les deux vaches fistulées utilisées ont été élevées en respectant la législation nationale sur l'expérimentation animale (certificat d'autorisation N°004495 du Ministère de l'Agriculture). Les animaux ont été nourris à volonté chaque matin, suivant un régime alimentaire classique, composé d'ensilage de maïs (64% du poids sec), de foin (6% du poids sec) et de concentrés (30% du poids sec). Le prélèvement de rumen a été réalisé juste avant un repas. Le contenu ruminal, prélevé en différentes localisations dans le rumen, a été homogénéisé manuellement avant que des échantillons de 30g soient congelés à l'azote liquide, puis d'être conservés à -80°C. Afin de limiter l'effet de variations individuelles sur la suite du protocole, un mélange d'un échantillon de chacun des deux rumens a systématiquement été utilisé comme inoculum.

I.2. Termites

I.2.1 Choix des espèces

Afin d'augmenter les chances d'obtenir un consortium stable et actif sur lignocellulose à partir de bactéries intestinales de termites, quatre espèces différentes de termites ont été sélectionnées pour servir d'inoculum à une première expérience de fermentation. Les critères déterminant le choix des espèces ont été la composition de la flore (désirée à forte dominance bactérienne), le pH (légèrement acide, ou le moins basique possible), le régime alimentaire et la production d'acétate. Après discussion avec des spécialistes de l'IRD, les termites inférieurs, dont la flore intestinale est composée d'un mélange d'eucaryotes unicellulaires en symbioses complexes avec l'hôte et des bactéries, ont été écartés, leur flore étant probablement très difficile à stabiliser en réacteur. Dans les termites supérieurs, les termites champignonnistes ont également été écartés, la digestion de la lignocellulose étant assurée par une symbiose externe avec un champignon. Dans les espèces restantes, et en fonction de la disponibilité des espèces dans les élevages de l'IRD, le choix s'est porté sur *Microcerotermes*

parvus, *Nasutitermes ephratae*, *Nasutitermes lujae*, et *Termes hospes*. *Microcerotermes* est un genre contenant une vingtaine d'espèces arboricoles, se nourrissant principalement de bois mort (régime alimentaire de type 2). Le genre *Nasutitermes* contient un très grand nombre d'espèces, la plupart arboricoles. Il est également d'un régime alimentaire de type 2. De manière intéressante, la poche P3 (la plus importante) de son intestin présente un pH légèrement acide, alors qu'elle est très basique chez la plupart des insectes se nourrissant de bois. Enfin, *Termes* est un genre humivore (type 3) se nourrissant de bois décomposé ou de sol à forte teneur organique.

I.2.2 Récolte et conditionnement

Les termites collectés proviennent tous de colonies élevées à l'IRD de Bondy, en chambre contrôlée (27°C et 60% d'humidité relative). Pour chaque réacteur, cinq cent ouvriers (les soldats et juvéniles ont été écartés) s'étant aventurés sur un morceau de bois frais déposé sur le nid ont été collectés aléatoirement (les soldats et juvéniles ont été écartés). Après une anesthésie sur glace, les intestins ont été disséqués à l'aide de pinces fines stérilisées. Une fois retirés, ils ont été immédiatement transférés dans une solution saline maintenue sur glace. Tous les 250 intestins collectés, l'ensemble a été congelé à -20°C, puis expédié au laboratoire sur glace carbonique avant d'y être stocké à -80°C. Chaque espèce a été testée en deux réplicats de réacteur, la collecte de termites ayant été réalisé à trois mois d'intervalle. À des fins de séquençage et de quantification du nombre de copies 16S par intestin, 2x20 intestins supplémentaires ont été collectés. Enfin, des groupes de 10 ouvriers ont été pesés pour estimer la masse de l'inoculum.

Espèce	Poids humide (mg pour 20 ouvriers)
Microcerotermes parvus	61,7
Nasutitermes ephratae	64,8
Nasutitermes lujae	48,8
Termes hospes	24,2

II. Dispositifs expérimentaux de culture en bioréacteur

II.1. Bioréacteurs du screening termites

II.1.1 Dispositif expérimental

Afin de favoriser l'établissement des communautés et du fait de la petite taille des inocula, des réacteurs de petit volume (400mL) ont été inoculés avec les intestins de termites. Les réacteurs utilisés ont été des Applikon MiniBio 500, utilisés en conditions anaérobies et après stérilisation à l'autoclave. Les échantillons d'intestin ont été centrifugés (7197g, 10 min, 4°C) et la solution saline éliminée. 500 intestins centrifugés ont alors servi à inoculer 400 mL de milieu minéral (MM, composition détaillée ci-dessous) additionnés de 20g.L^{-1} de paille de blé comme seule source de carbone (paille broyée à 2mm de variété Koreli, récoltée sur une parcelle INRA de Boissy-le-Repos, 51210 France en août 2011). Afin d'éviter la présence de bactéries endogènes du substrat, la paille a été stérilisée à l'autoclave (120°C, 20 min et 1,2 bars en présence d'eau). Une fois les réacteurs inoculés, les conditions anaérobies ont été obtenues par injection d'un flux d'azote avant fermeture hermétique du réacteur. La température a été maintenue à 35°C et le pH à 6,15 par ajout d'une solution de soude à 2M NaOH, sous agitation constante à 400 rpm. La surpression due à la production de gaz a été libérée lors des prélèvements gaz.

Composition du milieu minéral MM (pour un litre) :

KH_2PO_4 : 0,45 g ; K_2HPO_4 : 0,45 g ; NH_4Cl : 0,4 g ; NaCl : 0,9 g ; $MgCl_2.6H_2O$: 0,15 g ; $CaCl_2.2H_2O$: 0,09g ; 250µL de solution de vitamine V7 (Pfennig and Trüper, 1992) ; 1mL d'oligo-éléments contenant pour un litre : H_3BO_3 : 300 mg ; $FeSO_4.7H_2O$: 1,1 g ; $CoCl_2.6H_2O$: 190 mg ; $MnCl_2.4H_2O$: 50 mg ; $ZnCl_2$: 42 mg ; $NiCl_2.6H_2O$: 24 mg ; $NaMoO_4.2H_2O$: 18 mg ; $CuCl_2.2H_2O$: 2 mg. Solution stérilisée par filtration (0,2 µm)

II.1.2 Paramètres suivis

La taille du dispositif expérimental ne permettant pas de prélèvement du milieu totale, seules des mesures de gaz et de liquide ont été possibles. La production de gaz a été mesurée quotidiennement par des mesures de pression et des prélèvements de gaz analysés par chromatographie gazeuse (HP 5890 équipée d'une colonne HAYSEP D de 5Å avec argon en gaz porteur à 100 mL.min^{-1}). Les prélèvements liquides ont permis de suivre l'accumulation

des acides gras volatils au cours du temps. La dégradation (estimée à partir des matières volatiles résiduelles) n'a pu être estimée qu'au point final une fois le réacteur arrêté.

II.2. Bioréacteurs d'enrichissement rumen et termite

II.2.1 Dispositif expérimental

Les enrichissements ont été réalisés en réacteurs batch successifs (SBR) en utilisant des BIOSTAT 2L A+ (Sartorius, Germany) et des conditions identiques à celles décrites en partie II.1.1 (anaérobie, 35°C, pH=6,15, agitation continue, même milieu de culture), à deux différences près : le gaz produit a été mesuré en ligne grâce à une sortie gaz connectée à un volumètre Ritter, et le substrat carboné (paille de blé à 20g.L^{-1}, origine et lot identique) n'a pas été stérilisé afin de favoriser la sélection d'un inoculum robuste face à l'invasion de bactéries endogènes du substrat.

Le premier cycle a été initié en inoculant le réacteur avec 40 grammes de rumen (poids humide) ou 200 mL de la première culture de bactéries de *Nasutitermes ephratae*. La culture est poursuivie jusqu'à atteindre un plateau de dégradation (7 à 10 jours). À la fin de chaque cycle, le cycle suivant est inoculé avec 10% en volume du cycle précédent. Une fois les capacités et la communauté stabilisées, un dernier cycle a permis de produire des séries d'échantillons de 200 mL, congelés dans l'azote liquide et conservés à -80°C qui ont servi d'inoculum aux cinétiques de caractérisation.

II.2.2 Paramètres suivis

Le volume de réacteur plus important ainsi que les tailles de connecteurs ont permis des prélèvements de réacteurs totaux en cours de culture, donnant accès aux valeurs de dégradation (mesure des matières volatiles résiduelles et sucres résiduels) au cours du temps, en plus des valeurs d'acides gras volatils mesurées sur la fraction liquide des prélèvements. La production de gaz a été mesurée en ligne par volumètre Ritter, sa composition par chromatographie gazeuse. Des prélèvements réguliers de réacteur ont également été effectués afin de caractériser la diversité bactérienne et sa stabilisation au cours de l'enrichissement.

II.2.3 Bioréacteurs paille stérile / non stérile (termites)

Afin d'étudier un possible effet des bactéries endogènes du substrat sur l'enrichissement, un deuxième enrichissement à partir de bactéries intestinales de termite a été réalisé exactement dans les mêmes conditions, mais avec une paille stérilisée par autoclave (substrat sec à 120°C, 20 min et 1,2 bars). Les conditions de culture et paramètres suivis ont été conservés entre enrichissement sur substrat stérile et enrichissement sur substrat non stérile.

II.3. Bioréacteurs de cinétiques de caractérisation

II.3.1 Dispositif expérimental

Les réacteurs et leur configuration sont les mêmes pour les cinétiques de caractérisation qu'en cours d'enrichissement. La seule différence avec un cycle d'enrichissement est la congélation de l'inoculum à la fin du dernier cycle. Chaque réacteur de caractérisation a été inoculé avec un échantillon de 200 mL, précédemment décongelé à 4°C pendant une nuit. Chaque cinétique de caractérisation a été réalisée en duplicat ou triplicat avec des prélèvements quotidiens ou biquotidiens pendant 15 jours.

II.3.2 Paramètres suivis

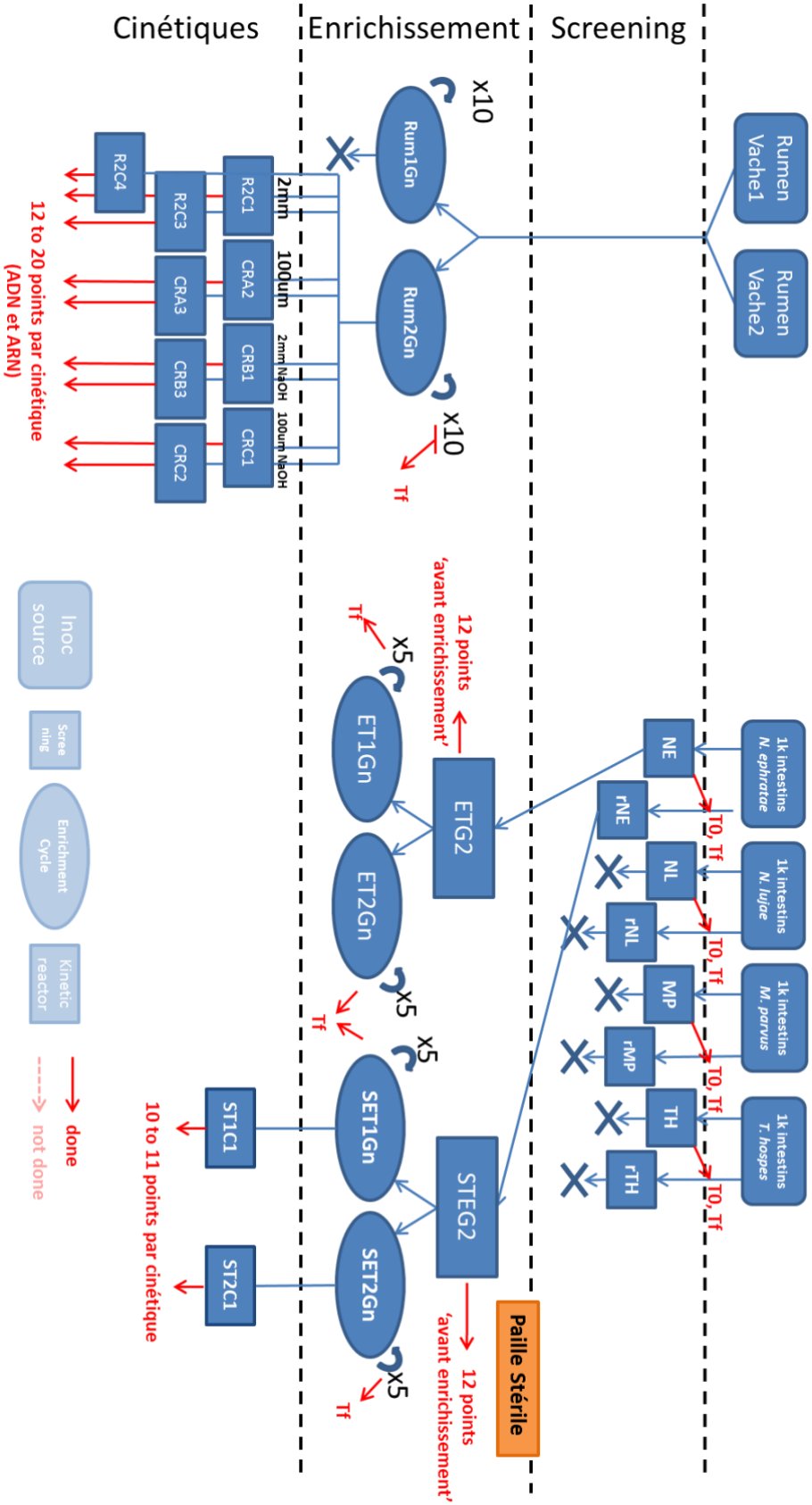
Des prélèvements quotidiens voire biquotidiens ont été réalisés pour caractériser les cinétiques de dégradation et de production, mais également les activités enzymatiques liées à la transformation de lignocellulose, les dynamiques de populations bactériennes et leur activité. Ainsi, la dégradation a été suivie par mesure des matières volatiles résiduelles et dosage des sucres résiduels mais également par spectroscopie infrarouge FT-IR. La production de gaz a été mesurée en ligne et sa composition mesurée ponctuellement par chromatographie gazeuse, et la production d'acides gras volatils a été suivie quotidiennement. Les activités CMCase et xylanase (représentatives respectivement des activités cellulase et hemicellulase) ont été suivies par prélèvements quotidiens. Enfin, la dynamique des communautés bactériennes et de leur activité, évaluée par séquençage de l'ADNr 16S et de l'ARNr 16S a été suivie tout au long des cinétiques de caractérisation.

II.3.3 Prétraitements

La paille de blé broyée à 2mm par un broyeur à couteaux (Retsch SM 100, Germany) représente la paille de référence (traitement A). À partir de celle-ci, trois pailles prétraitées ont été obtenues par imprégnation de soude (traitement B), broyage à 100 µm avec un broyeur à boulets (traitement C) et par imprégnation de soude pendant un broyage à 100µm (traitement D). Le prétraitement a été réalisé à l'INRA de Montpellier dans l'équipe d'A. Barakat selon la procédure décrite dans leurs travaux (Barakat et al., 2013, 2014).

II.3.4 Plan d'expériences

Le plan d'expérience global est le suivant :



III. Analyses macrocinétiques

III.1. Mesure des matières volatiles résiduelles

Les matières sèches (MS), matières volatiles (MV) et matières minérales (MM) sont mesurées à partir de prélèvements de 10 mL de réacteur total. Les échantillons sont centrifugés (7197 g, 4°C, 10 min) et rincés deux fois à l'eau distillée, avant d'être séchés 24h à 105°C. Les matières minérales sont mesurées par minéralisation des échantillons à 500°C pendant 2h, et les matières volatiles calculées par différence entre les pesées de matières sèches et matières minérales. La dégradation, reportée en pourcentage massique se rapporte aux matières initiales avant inoculation.

III.2. Dosage des acides gras volatils produits

L'accumulation d'acides gras volatils (AGV) a été déterminée à l'aide de prélèvements de la phase liquide du réacteur (arrêt d'agitation jusqu'à sédimentation et prélèvement par seringue). La composition et la concentration en acétate, propionate, isobutyrate, butyrate, isovalérate, valérate et hexanoate ont été déterminées en utilisant un chromatographe gazeux Varian 3900 équipé d'une colonne CB CP-Wax 58 (FFAP).

III.3. Dosage des sucres résiduels

Pour mesurer les sucres résiduels, des échantillons de 10 mL de réacteur sont séchés à 105°C et la masse de matière sèche correspondante mesurée. Une hydrolyse acide en deux étapes est alors réalisée. Pour la première étape, 40 mg de paille séchée sont incubés 1 h à 30°C dans 500 µL de H_2SO_4 concentré (72%). Pendant la deuxième étape, l'acide est dilué par ajout d'eau distillée pour atteindre une concentration de 10%, et incubé à 100°C pendant 90 min. La quantification des monomères de sucres libérés par l'hydrolyse est alors réalisée par chromatographie liquide à haute performance (HPLC) avec un Ultimate 300 Dionex équipé d'une colonne BioRad Aminex HPX87H. La séparation des produits est réalisée dans du H_2SO_4 à 5 mM, 40°C et un flux de 0,3 $mL.min^{-1}$.

III.4. Mesure des activités enzymatiques

Les activités enzymatiques ont été mesurées à partir de triplicats de prélèvement (5 mL). Les échantillons ont été centrifugés à 7197xg et 4°C pendant 10min pour séparer surnageant (enzymes extracellulaires) et culot (enzymes intracellulaires ou liées). Le culot a été suspendu dans 6 mL de tampon acétate (50 mM pH=6), transféré sur glace et passé aux ultrasons 4x20s à 60W (Bandelin Sonoplus HD 2070, sonde MS73). Les activités xylanase et CMCase ont été mesurées en utilisant 1% w/v de xylan de bouleau (Sigma) et 1% w/v de carboxymethylcellulose (CMC, Sigma) dissous en solution tampon. Elles ont été calculées à partir de la quantité de sucres réducteurs libérés pendant une incubation à 35°C pendant 1 heure (xylanase) ou 4 h (CMCase). Les sucres libérés ont été mesurés par la méthode DNS (acide dinitrosalicyclique) à l'aide d'un spectrophotomètre UV (Multiskan Ascent, Thermo Scientific).

IV. Outils moléculaires

IV.1. Méthode de co-extraction ADN/ARN

Les échantillons de 1,5 mL prélevés ont été centrifugés à 13 000xg, 4°C pendant 5 min. Le surnageant a été retiré et le culot congelé dans l'azote liquide avant d'être stocké à -80°C jusqu'à extraction. Une co-extraction permettant de récupérer ADN et ARN a été réalisée à l'aide du kit PowerMicrobiome RNA Isolation kit (MoBio Laboratories Inc. Carlsbad) en suivant le protocole constructeur mais en omettant les étapes optionnelles de traitement DNAse. La lyse cellulaire a été effectuée avec un FastPrep (MP Biomedicals, 2x30 secondes à $4ms^{-1}$). ADN et ARN ont alors été séparés en utilisant un AllPrep DNA/RNA MiniKit (Qiagen) en suivant le protocole constructeur. La qualité de l'extraction a été vérifiée par électrophorèse sur un gel d'agarose 1% et les quantités d'ADN et ARN purifiés obtenues ont été mesurées par absorbance à 260 et 280nm (Nanodrop 1000, ThermoScientific).

Toute trace potentielle d'ADN contaminant dans les échantillons d'ARN purifié a été éliminée en utilisant un kit TURBO DNA-free (Ambion) en suivant le protocole constructeur. La qualité de l'ARN obtenue a été vérifiée avec un BioAnalyzer 2100 (Agilent Technologies) et une puce RNA Pico 6000 Chip. Les ARN ont alors été rétro-transcrits en utilisant un kit M-MLV Reverse Transcriptase (Promega) et des hexamères aléatoires comme amorces, en suivant les instructions du constructeur.

IV.2. Analyses par amplification de l'ADN 16S

IV.2.1 Dosage des copies 16S par qPCR

Le nombre de copies du gène de l'ADNr 16S a été déterminé par qPCR à partir des échantillons d'ADN purifié en utilisant un Realplex Mastercycler (Eppendorf, Montesson, France). Les mesures ont été réalisées en triplicat et dilutions successives pour correspondre à la gamme étalon (plasmide pEX-A contenant la séquence cible d'*E. coli* (Eurofins MWG Operon). La réaction de qPCR a été réalisée dans 25 µL en suivant les instructions du constructeur pour une solution contenant 12,5 µL de MasterMix (Invitrogen, Eugen, USA) avec les primers BAC338F et BAC805R (250 nM chacun), la sonde TaqMan BAC516F (100 nM) et une quantité d'ADN cible comprise entre 10 et 100 ng. Les cycles PCR ont été fixés comme décrit par Yu et al. (2005) : 20 secondes à 95°C suivies de 40 cycles de 15 secondes à 90°C et 1min à 60°C. Un contrôle négatif sans ADN a subi la même procédure pour s'assurer de l'absence de contamination et seules les dilutions ne présentant pas d'inhibition PCR ont été utilisées pour calculer le nombre de copies.

IV.2.2 Préparation des librairies MiSeq Illumina

La diversité bactérienne et son activité ont été étudiées par séquençage de l'ADNr 16S et de l'ARNr 16S en ciblant les régions hypervariables V3 et V4 et en utilisant un séquenceur Illumina MiSeq à la plateforme de séquençage GenoToul Genomics and Transcriptomics (GeT, Auzeville, France). Le séquençage a été réalisé en pair-end.

.IV.2.2.1 PCR1 de préparation des librairies (amplification V3-V4)

La région V3-V4 du gène de l'ARNr 16S a été amplifiée par PCR en utilisant les amorces bactériennes 343F et 784R, modifiées pour ajouter des adaptateurs de la PCR1 :

343F=5'-CTT-TCC-CTA-CAC-GAC-GCT-CTT-CCG-ATC-TAC-GGR-AGG-CAG-CAG-3'
784R=5'-GGA-GTT-CAG-ACG-TGT-GCT-CTT-CCG-ATC-TTA-CCA-GGG-TAT-CTA-ATC-CT-3'

La PCR1 a été réalisée dans 50 µL de milieu réactionnel contenant 1X de tampon et 2,5U de MTP Taq DNA Polymerase (Sigma), 0,2 µM de chaque dNTP, 0,5 µM de chaque amorce et 2ng d'ADN purifié ou 1 µL d'ADNc dilué au vingtième (~0,1 ng d'ADNc). L'amplification par PCR a été réalisée en 30 cycles de (1min : 94°C ; 1min : 65°C ; 1min : 70°C). Les

produits obtenus ont alors été purifiés en utilisant des billes magnétiques avant d'être quantifiés par un spectrophotomètre Nanodrop 1000.

.IV.2.2.2 PCR2 de préparation des librairies (ajout index et adaptateurs)

Une deuxième réaction de PCR a alors été réalisée pour ajouter les adaptateurs de séquençage ainsi qu'un index unique pour chaque échantillon, en utilisant les amorces suivantes :

FP2=5'-AAT-GAT-ACG-GCG-ACC-ACC-GAG-ATC-TAC-ACT-CTT-TCC-CTA-CAC-GAC-3'
RP2=5'-CAA-GCA-GAA-GAC-GGC-ATA-CGA-GAT-**index**-GTG-ACT-GGA-GTT-CAG-ACG-TGT-3'

La PCR a été réalisée dans 50 µL de milieu réactionnel contenant 1X de tampon et 2,5U de MTP Taq DNA Polymerase (Sigma), 0,2 µM de chaque dNTP, 0,5 µM de chaque amorce et 15ng du produit PCR purifié à l'étape précédent. L'amplification par PCR a été réalisée en 12 cycles de (1min : 94°C ; 1min : 65°C ; 1min : 72°C). Les produits obtenus ont alors été purifiés en utilisant des billes magnétiques avant d'être quantifiés par un spectrophotomètre Nanodrop 1000. La qualité et la mesure de quantité d'un dixième des échantillons choisis aléatoirement ont été vérifiées en utilisant un BioAnalyzer 2100 et un kit High Sensivity DNA Analysis (Agilent). La librairie a été préparée par mélange équimolaire à partir de tous les échantillons à séquencer, et dosée par qPCR.

.IV.2.2.3 Lancement du séquençage sur MiSeq

Le séquençage a été réalisé en chargeant 1 mL de librairie à 7 pM sur une cartouche Illumina MiSeq et un kit de réactifs v2 ou v3 selon les fois. Les données répondant aux filtres de qualité interne au séquenceur ont été vérifiées par la plateforme GeT. Elles ont alors été démultiplexées (deux fichiers de sortie par échantillon) puis les lectures jumelles ('pair-end reads') ont été jointes en une seule, plus longue ('contig') en utilisant le logiciel Flash v1.2.6 (Magoč and Salzberg, 2011) avec un recouvrement minimum de 110 paires de base et un ratio de mismatch (non-appariement) maximum de 0,1.

V. Traitement des données de séquençage (pipeline IPS)

Le traitement des données a été réalisé principalement à partir de fonctions Mothur (Schloss et al., 2009), ajoutées de scripts « maison » en Bash et Python. L'architecture de base est celle de la procédure standard d'utilisation décrite par Kozich et al. (2013) avec quelques modifications, testées et discutées dans le chapitre III. Toutes ces étapes sont enchaînées automatiquement dans un pipeline écrit en Bash (IPS), ne demandant en entrée que les fichiers fastq des échantillons à traiter, une liste de ces échantillons, le seuil à appliquer pour filtrer les OTUs rares, et la taille de sous-échantillonnage désirée.

V.1. Formatage et création d'un fichier fasta unique

À partir des fichiers fastq (un par échantillon) fournis par la plateforme de séquençage et d'une liste d'échantillons associant nom de fichier et nom d'échantillon, un script Python permet de fusionner tous les fichiers fastq en un unique fichier associé à un fichier d'origines (ou fichier .groups) permettant de conserver l'attribution de chaque séquence du fasta à un échantillon. Le fichier fastq obtenu est alors converti au format fasta, l'information de qualité contenue dans le fastq n'étant pas utilisée par la suite (elle était utilisée par Flash). Un rapport de qualité des séquences est tout de même généré pour visualisation de la qualité des données de séquençage.

V.2. Nettoyage des données, pre-clustering, filtres de bruit et dé-chimérisation

Dans une deuxième étape réalisée quasiment uniquement à partir de fonctions Mothur et de quelques articulations en Bash, les amorces sont retirées à chaque séquence, et les séquences présentant un ou plusieurs mismatchs avec la séquence des amorces sont éliminées (fonction *trim.seqs*). Les séquences correspondantes sont également retirées du fichier d'origines (.groups) à l'aide d'un script Bash et de la fonction *remove.seqs* de Mothur. Les séquences porteuses de N (base indéterminée), d'homopolymères de longueur supérieure à 12, ou ne présentant pas la longueur attendue sont alors filtrées (fonction *screen.seqs*) et retirées simultanément du fichier d'origines. Le fichier est alors dérépliqué, ce qui consiste à ne garder qu'un exemplaire de chaque séquence (« séquence unique ») présente dans le fichier (fonction *unique.seqs*). L'information de leur effectif, ou abondance, est stockée dans un fichier contingent. De cette façon, la taille du fichier fasta est grandement réduite et les opérations qui y sont appliquées, plus rapides et moins coûteuses en termes de mémoire et de

puissance de calcul. La dernière étape de filtre qualité consiste à aligner toutes les séquences uniques contre une base multi-alignée (dans notre cas LTP115, découpée sur V3-V4) et à éliminer toutes les séquences ne s'alignant pas aux positions attendues (fonctions *align.seqs*, *screen.seqs* et *filter.seqs*). Un filtre de toute les séquences singleton, c'est-à-dire de toutes les séquences qui ne sont présentes qu'une seule fois dans tout le jeu de données, peut être envisagé à cette étape (méthode nommée SRF1, voir chapitre III) et peut être réalisé avec la fonction *split.abund* de Mothur.

L'étape suivante, appelée *preclustering* (fonction *pre.cluster*) consiste à corriger les petites erreurs de séquençage. Celles-ci, des substitutions de bases, sont supposées intervenir aléatoirement et être plus rares que la séquence vraie. Si une séquence très rare ne présente que quelques différences (paramètre *d*) avec une séquence très abondante, on considère alors qu'elle est issue d'une erreur de lecture de la séquence abondante. On l'élimine alors, en ajoutant son effectif à l'effectif de la séquence qui est conservée. Le preclustering est réalisé avec un paramétrage de *d*=5. Un filtre des séquences singleton peut être envisagé à cette étape (méthode SRF2, voire chapitre III).

La détection de chimères est ensuite réalisée en utilisant la version de Uchime implantée dans Mothur, *chimera.uchime* et mode référence interne. Les séquences identifiées comme chimériques sont éliminées via la fonction *remove.seq* de Mothur. Une étape de sous-échantillonnage (fonction *sub.sample*) par tirage aléatoire peut être appliquée à cette étape (méthode SUB du chapitre III) pour réduire le nombre de séquences à traiter aux étapes ultérieures.

V.3. Assignation taxonomique des séquences

L'étape de classification (ou affiliation taxonomique) des séquences est alors réalisée sur chaque séquence unique de jeu de donnée, en utilisant la fonction *classify.seqs* et la méthode Wang. La base de donnée utilisée n'est pas toujours la même et dépend des échantillons utilisés. Une fusion de LTP et de DictDB, base de données taxonomique spécialisée dans les bactéries associées à des insectes a été utilisée dans la plupart des cas (voir matériel et méthode chaque chapitre). Le seuil de certitude pour qu'une taxonomie soit conservée est fixé à 90%.

V.4. Calcul des distances et clustering

Le clustering est l'étape clé, et la plus coûteuse en mémoire et puissance de calcul du traitement de données. Il est de plus non parallélisable, ce qui peut le rendre extrêmement long, et est le point critique lors du traitement de très gros jeux de données. Il consiste à regrouper en « clusters » nommés OTUs (Operational Taxonomic Units) les séquences se ressemblant suffisamment, et à les séparer au-delà d'un seuil (fixé à 3%). Il commence par le calcul de distances entre séquences (étape parallélisable) à l'aide de la fonction *dist.seqs*, qui produit une matrice triangulaire de distance entre toutes les séquences. Sa taille correspond donc au demi-carré du nombre de séquences dans le jeu de données, et celle-ci doit être chargée en mémoire pour initier le clustering (fonction *cluster*). Il est donc nécessaire pour traiter de gros jeux de données, de réduire ce nombre de séquences, par des approches de sous-échantillonnage ou de filtres (voir chapitre III). Les OTUs obtenus peuvent être filtrés par abondance (les OTUs trop rares sont éliminés) avec la fonction *split.abund*, et le nombre de séquence par échantillon peut être normalisé par tirage aléatoire (*sub.sample*).

V.5. Assignation taxonomique des OTUs et calcul de distance entre OTUs

L'assignation taxonomique est réalisée à l'aide de l'assignation de chaque séquence déterminée précédemment. La taxonomie d'un OTU est la taxonomie majoritaire des séquences qui le composent (seuil fixé à 90%). La séquence majoritaire de chaque OTU est choisie comme séquence représentative (fonction *get.oturep*), et les séquences représentatives servent à construire un arbre phylogénétique (scripts Bash et logiciels ClustalO et raxmlHPC).

V.6. Résultats, sorties graphiques et import R

Les sorties indispensables sont les tables d'abondance (effectif de chaque OTU pour chaque échantillon) et la taxonomie correspondante. Celles-ci sont générées par la fonction *make.shared.* Des courbes de raréfaction (détection de nouveaux OTUs en fonction de la profondeur de séquençage) peuvent être produites à l'aide de la fonction *rarefaction.single.* Une sortie permettant de vérifier le bon déroulement du traitement de données est générée par un script Python et permet de suivre à chaque étape du pipeline le nombre de séquences par échantillon, afin de détecter d'éventuelles anomalies lors du traitement. Un fichier de visualisation Krona est également généré à l'aide de scripts Python et Perl.

L'analyse des données se fait en grande partie sous R, principalement avec les package Phyloseq (McMurdie and Holmes, 2013) et mixOmics (Cao et al., 2009). La table d'abondance, le fichier de taxonomie et l'arbre phylogénétique sont importés et formatés par des scripts R pour être compatibles avec les formats d'entrée de ces deux package.

VI. Analyse comparatives des dynamiques de communautés bactériennes

À partir des données traitées par le pipeline IPS et importées dans R aux formats Phyloseq et mixOmics, ces deux packages ont été utilisés pour comparer les échantillons entre eux et analyser les dynamiques de communautés à l'aide de différents outils. Le format Phyloseq permet de créer des « objets phyloseq » qui rassemblent les différents éléments (table d'abondance, taxonomie, arbre phylogénétique et métadonnées) en un seul objet plus simple d'utilisation. Le package mixOmics lui n'a pas de format particulier mais permet à partir de tableaux d'utiliser de nombreuses méthodes statistiques.

VI.1. Calcul des indices de diversité

Les indices de diversité (richesse, indices de Shannon et de Simpson) ont été calculés à partir des abondances en OTUs de chaque échantillon à l'aide de Phyloseq (fonction *plot_richness*). Celle-ci permet également à l'aide des métadonnées (origine de l'échantillon, inoculum de départ, temps de prélèvement, etc.) de calculer ces indices sur un groupe d'échantillon, ou à l'aide de la fonction *tax_glom*, de les calculer à des niveaux taxonomiques supérieurs à l'OTU. Phyloseq permet de tracer directement des figures, modifiables en syntaxe *ggplot*, ou de récupérer les valeurs des indices calculées pour les traiter en dehors de R.

VI.2. Statistiques exploratoires

VI.2.1 Analyses de projection de données (PCA et PLS)

Les analyses en composante principales ont été réalisées en utilisant les fonctions *pca* (analyse en composante principale ou PCA en anglais) et *spca* (sparse-PCA, ou sPCA, permettant de réduire le nombre de variables pour ne conserver que les plus significatives d'une composante) du package R mixOmics (Cao et al., 2009). Le nombre de composantes choisi, ainsi que le nombre de variables à conserver dans chaque composante pour la sPCA a été calibré pour obtenir une variance expliquée suffisante et une bonne séparation des points

analysés, en fonction des données et du niveau d'analyse (OTU, genre, classe ou phylum). Les graphes ont été obtenus à l'aide des fonctions *plotIndiv* et *plotVar*, avec un seuil de corrélation de 0,5 pour les variables.

À l'aide des métadonnées de chaque échantillon, formatées en une matrice complémentaire de la table d'abondance, les analyses de régression des moindres carrés partiels (ou PLS, permettant une analyse croisée des données contenues dans deux tables) ont été réalisées à l'aide des fonctions *pls* et *spls* (sparse-PLS, permettant également de réduire le nombre de variables dans chaque composante) de mixOmics, utilisées en mode régression. Le nombre de composantes à considérer a été calibré à l'aide de la fonction *perf* pour chaque jeu de données.

VI.2.2 Méthodes de projections de distances

Les distances de Bray-Curtis, distances Unifrac et Unifrac pondéré (weighted-Unifrac) ont été calculées des arbres calculés comme décrits précédemment (partie V.4) et de la fonction *distance* du package Phyloseq. L'organisation des échantillons entre eux en fonction de ces distances a été étudiée avec les méthodes de projection de Principle Coordinate Analysis (ou PCoA) et Non-Metric Dimensional Scaling (NMDS), utilisées par l'intermédiaire de la fonction *ordinate* de Phyloseq. Les graphes ont été organisés à l'aide des métadonnées incorporées dans l'objet Phyloseq à l'aide des fonctions de *ggplot*.

VI.3. Statistiques de test et classification supervisée

VI.3.1 ANOVA / PERMANOVA

La significativité de différences observées pour un paramètre (abondance d'un OTU ou métadonnées) a été évaluée par ANOVA (Analysis Of Variance) à l'aide des fonctions *anova* et *lm* de R. La significativité de regroupements de points observés avec les méthodes de statistiques exploratoires a été évaluée par PERMANOVA (Permutational Multivariate ANOVA) en utilisant les matrices de distances Bray-Curtis ou Unifrac et Unifrac pondéré, et la fonction *adonis* du package R vegan.

VI.3.2 sPLS-DA

L'identification de variables permettant la séparation voulue (classification supervisée) a été réalisée à l'aide de la fonction *splsda* de mixOmics, permettant d'appliquer une approche de sPLS en utilisant le critère de séparation comme deuxième matrice. Le nombre de composantes a été calibré à l'aide de la fonction *perf*. Les graphes ont été obtenus à l'aide des fonctions *plotIndiv* et *plotVar*, avec un seuil de corrélation de 0,5 pour les variables.

CHAPITRE III

VALIDATION DES METHODES DE TRAITEMENT DE DONNEES DE SEQUENCAGE

CHAPITRE III : VALIDATION DES METHODES DE TRAITEMENT DE DONNEES DE SEQUENCAGE

I. Introduction

L'étude des communautés microbiennes a beaucoup évolué avec les nouvelles technologies de séquençage. Avec le MiSeq Illumina, il est aujourd'hui possible de séquencer simultanément plusieurs centaines d'échantillons tout en atteignant une profondeur de plusieurs dizaines de milliers de séquences pour chaque échantillon. Les outils classiques de traitement des données, conçus alors que les données étaient de taille plus faible, rencontrent alors de grosses difficultés pour en gérer une quantité si importante. Différentes stratégies peuvent donc être adoptées : développer, tester et valider de nouveaux outils capables de traiter de très gros jeux de données, ou travailler avec des outils déjà validés mais réduire la taille des données, notamment en éliminant les erreurs de séquençage.

En effet, les nouvelles technologies de séquençage ne sont pas exemptes d'erreurs, qui ont pour principal effet d'augmenter artificiellement la diversité en générant de faux taxons (Kunin et al., 2010). Ceux-ci, en plus de polluer les données finales, ont par leur nombre élevé un impact extrêmement fort sur les performances et les capacités des outils de traitement. L'élimination des séquences porteuses d'erreur est donc un enjeu majeur du traitement de données issues de séquençage. Parallèlement aux outils complexes dédiés à leur détection, il existe un critère très simple permettant de suspecter la présence d'erreurs de séquençage : l'abondance. En effet, après plusieurs cycles de PCR lors de la préparation des librairies, une séquence représentée une unique fois sur les dix millions du jeu de données complet a de grandes chances de présenter des erreurs. Par ailleurs, si elle n'est pas regroupée avec d'autres séquences au sein d'un OTU, elle sera éliminée dans la majorité des pipelines d'analyse, qui filtrent les OTU constitués d'une seule séquence. La filtration, en amont, des séquences singleton, a déjà été proposés par d'autres équipes (Meeus et al., 2015) et est proposée par défaut dans le logiciel UPARSE (Edgar, 2013), et permet des gains de temps significatifs. Cependant, son impact sur les résultats et les relations entre communautés bactériennes n'a jusqu'ici pas été étudié.

Ce chapitre décrit donc la validation d'une méthode de filtration par une approche reposant sur la comparaison de résultats obtenus avec différentes méthodes de filtre, en utilisant des données simulées, mais aussi réelles (communautés artificielles, réelles, et très gros jeux de données).

II. Analyse de gros jeux de données 16S Illumina : impact d'un filtre de séquence singleton sur la description des communautés microbiennes.

Ce chapitre sera prochainement soumis à « Molecular Ecology Resources », sous le titre :

Analysis of large 16S rRNA Illumina datasets: impact of singleton read filtering on microbial community description.

Lucas Auer[1,2,3], Mahendra Mariadassou[4], Michael O'Donohue[1,2,3], Christophe Klopp[5], Guillermina Hernandez-Raquet[1,2,3]*

[1)] Université de Toulouse, INSA, UPS, LISBP, 135 Avenue de Rangueil, F-31077 Toulouse, France
[2)] INRA, UMR792 Ingénierie des Systèmes Biologiques et des Procédés, F-31400 Toulouse, France
[3)] CNRS, UMR5504, F-31400 Toulouse, France
[4)] MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[5)] INRA, UR 875 GenoToul Bioinformatics Facility, Toulouse, France

II.1. Abstract

Diversity of bacterial communities is commonly studied by sequencing the variable regions of the 16S rRNA gene. Advances in sequencing technologies, especially the use of Illumina technology, provide access to more and more sequences at lower costs. However, the production of such large data sets also increases required calculation time and disk space. The data sets also accumulate technical bias and sequencing noise. To overcome both calculation and noise issues, recent studies and tools suggested removing all single reads, considering them as noise. This procedure increased the accuracy of OTU number predictions and reduced computational load. However, its effect on α and β diversity has been poorly studied.

In this study we test the effect of singleton read filtering (SRF) on community composition description using simulated datasets, synthetic and real communities. Scalability

to large datasets is also assessed using a complete MiSeq run. We show that SRF drastically reduces computational time while preserving compositions: the differences in compositions between SRF and standard procedures are much smaller than the intrinsic variability of technical and biological replicates.

II.2. Introduction

Microbial communities are ubiquitous within the environment where they drive major biogeochemical processes, including the recycling of carbon and nitrogen. Microbial communities are also used as biocatalytic systems to perform bioremediation of waste streams and bioconversion of other feedstocks into useful products. The optimal exploitation of these microbial resources requires knowledge of the community composition in addition to its metabolic functions and ecology. The discovery of universal genomic markers such as 16S rRNA gene and the recent sequencing technology progresses opened new perspectives for exploring and exploiting these microbial communities.

So far the diversity of microbial communities has mainly been studied through the analysis of 16S rRNA genes using next generation sequencing (NGS) technologies. Typically NGS technologies such as that developed by Roche (Roche 454 GS XLR instrument) generate, at reasonable cost and in a single run, one million DNA sequences, each with an approximate length of 400 nucleotides (nt), whereas the Illumina MiSeq technology can procure up to 20 million pair-end sequences of 250 nt per run. Once obtained, the sequences are usually clustered into Operational Taxonomic Units (OTUs) defined on the basis of 16S rRNA gene sequence divergence, with 3% divergence being a commonly accepted threshold, because this approximately corresponds to the species level (Hugenholtz et al., 1998).

Despite the power of NGS and the progress that these technologies have procured, the data that is generated is imperfect, being subject to different types of errors, including those inherent to PCR amplification (substitutions and chimera formation) and sequencing-specific bias that is characteristic of each sequencing technology or platform. Species affiliation errors are predominately caused by base substitutions, although base deletions, low-quality reads, variable read lengths, non-target amplification and inappropriate clustering are also known sources of error (Huse et al., 2010). The consequence of sequencing errors is inflated diversity estimates that translate into the creation of false taxa (Kunin et al., 2010). Undetected chimeric sequences, caused by the hybridization of DNA fragments from different species,

also reduce the reliability of 16S rRNA sequence-based community compositions (González et al., 2005; Jumpponen and Johnson, 2005). Together, these different errors generate a high number of isolated sequences only detected once, hereafter called singleton reads, which lead to overestimations of actual community diversity. Indeed, previous studies have shown that most singletons in pyrosequencing data resulted from DNA sequencing errors creating false OTUs (Tedersoo et al., 2010).

Over the past decade, tools for the detection and correction of sequencing errors have been developed and implemented in bioinformatics workflows. These mostly operate well with 454 sequencing technology, but generally display poor ability to scale to Illumina technology, which has a lower error rate, but up to 10 times higher throughput. Therefore, to process a complete Illumina run, dedicated computing facilities were required. New fast and less power-consuming tools, such as UPARSE (Edgar, 2013) are already available and others are under development, but currently none of them are completely consensual for the microbial ecologist community. While one solution at await the development of new, frugal and more accurate tools, another more pragmatic option is to reduce the size of the datasets by subsampling or by applying stringent filters that discard most error-prone reads. However, in this case it is essential to assess the impact of dataset size reduction on the accuracy of community diversity analysis.

Previous studies showing that an increase in sequencing depth is not a prerequisite to establish inter-sample relationships, or to estimate β-diversity, have provided the basis for subsampling strategies (Caporaso et al., 2011). However, such approaches may not be sufficient for very large projects, and could sometimes be wasteful or not accepted (McMurdie and Holmes, 2014). Nevertheless, assuming that sequencing depth is not an issue, large projects could be tackled by deleting all potentially erroneous sequences, defining these by examining sequence abundancy. After several PCR amplification cycles, a sequence that is found only once in ten million reads is likely to contain errors, and even if the sequence is genuine, its low frequency means that it will have a low impact on downstream analysis. Moreover, singleton OTUs, which are necessarily the result of singleton reads, are frequently removed from analyses (Medinger et al., 2010; Tedersoo et al., 2010), thus their systematic removal could be a good strategy to improve speed and accuracy. The use of this approach is discussed from a theoretical standpoint in the USEARCH manual (Edgar, 2010), is implemented as a default parameter in UPARSE (Edgar, 2013), and has already been proposed in Mothur (Meeus et al., 2015). However, so far the legitimacy of the systematic

removal of singleton reads has not been demonstrated and the quantitative effect of singleton reads in sequence-based species detection and its consequences on community composition and diversity estimates have not been thoroughly examined. To address this shortcoming, we have assessed the effect of a singleton read filtering (SRF) approach on α- and β-diversity estimates and distance calculations using simulated Illumina datasets. Moreover, we have examined the SRF effect on the perceived diversity of well-defined mock and real microbial communities. With these data, the SRF approach is not only faster, but also more accurate, reducing the number of false OTUs, while conserving inter-sample relationships. With very large datasets (20 million reads), SRF also enables the use of common hierarchical clustering tools such as Mothur, one of the most popular, but also most computationally expensive clustering algorithm.

II.3. Results

The analysis of sequencing data specific for the V3-V4 16S rRNA gene region was performed using Mothur v1.33.1 (Schloss et al., 2009), deploying four different strategies when they were compatible with available calculation capacities. The first strategy basically deployed the Mothur standard operating procedure (SOP). Hereafter, this is referred to as the reference method. The second approach involved a 20% subsampling step (SUB), and the two last strategies applied singleton read filters (SRF1 and SRF2). In all cases, data were trimmed, quality filtered and de-replicated before alignment and pre-clustering. This last step merges rare sequences with abundant ones when they present less than 5 nt differences, and thus corrects for small sequencing errors. Chimeras were then removed from the dataset and the curated reads were clustered using the average neighbor method, as implemented in Mothur. Prior to generating abundance tables, an additional filter for rare OTUs was applied to eliminate those with less than 20 sequences across all samples. This filter was considered as a detection threshold that approximately corresponds to 0.005 % of the total sequences, a value that has been proposed by recent studies (Bokulich et al., 2013). Subsampling (SUB) was done just before the clustering step and the singleton read filter was performed either before (SRF1), or after (SRF2) the pre-clustering step. The results procured by the four methods were analyzed in terms of total reads, observed OTUs and diversity indices, and the relationships between samples were analyzed using classical community ecology methods (e.g. Bray-Curtis or weighted-Unifrac distances and Principle Coordinate Analysis (PCoA) ordinations for visual exploration).

II.3.1 Simulated datasets

To study the impact of singleton read filtering, Illumina 16S rRNA V3-V4 region datasets (named SIM) were first generated *in silico* (see material and methods for details). The 100 sequences that maximize the phylogenetic diversity within the LTP database were selected. This corresponds to a favorable situation where species are as distinct as possible and 3% divergence clustering from perfect sequences recovers exactly 100 OTUs. The error rate profile was determined by sequencing the V3-V4 16S rRNA gene region of a pure culture of *Clostridium perfringens*, applying the MiSeq SOP method proposed by Mothur. The fraction of chimeras was set at 10% and the breakpoint position on the read followed a uniform distribution. Ten datasets containing 100 000 sequences each were simulated following a power law distribution, but with different species rank abundance. As the community composition and distribution are known for this SIM dataset, it was possible to compare the results obtained by the different methods to the actual ones. The simulated data also enabled us to quantify the fraction of remaining chimeras at each step of the analysis, information that cannot be accurately determined from mock NGS sequencing data or that of real microbial communities due to the presence of sequencing errors.

The total number of sequences, unique sequences, OTUs and chimeras at each step of SIM data processing are detailed in Table 1. SRF1, applied after the pre-clustering step, discarded 35% of the total sequences, reducing the number of unique sequences by 99.9%. In contrast to SRF1, SRF2, applied after the pre-clustering step, discards only 4% of the total sequences, corresponding to a 98% reduction of unique sequences. After chimera detection, taking SOP as a reference (12,335 unique sequences), the size of the clustering input was reduced 73 and 5.4 fold with SRF1 (169 unique sequences) and SRF2 (2,292 unique sequences), respectively. Due to the quadratic relationship between input size, clustering time and required memory, the singleton filter approach drastically improved Mothur capacities. With simulated datasets, clustering time was strongly reduced, from 70 min for SOP to 55s for SRF1, resulting in 6,280 and 126 OTUs. The calculation time for SUB and SFR2 were 3 and 6 min and resulted in 2,659 and 781 OTUs respectively. The last step of rare OTU filter, set at 0.002% of the initial sequences, brought the total number of observed OTUs of the ten datasets to a value of 100, 103, 138 and 144 OTUs when processed by SRF1, SUB, SRF2 and SOP, respectively. Since OTUs do not all appear in all datasets, the actual number of OTUs per dataset varied from 100 to 110 (Table 1). SRF1 was thus the faster, more accurate method,

at least concerning the expected number of OTUs. Only 6 errors in taxonomic assignation were observed, 3 at the species level (2 unclassified and 1 error) and 3 at the genus level (1 unclassified and 2 errors).

Table 1: Summary of SIM-data processing by the four methods. For each method, the total number of sequences (nSeq), unique sequences (nUnique) and chimeras are indicated. Processing times are indicated for computationally intense steps. For each method, the observed number of OTUs after clustering is indicated (nOTU), before and after applying the filter for rare sequences (RareFilter).

Step		**Method**							
		SOP	*Chimeras*	**SUB**	*Chimeras*	**SRF1**	*Chimeras*	**SRF2**	*Chimeras*
Filtering	nSeq	987 494	*98 017*	987 494	*98 017*	987 494	*98 017*	987 494	*98 017 (9.9%)*
	nUnique	427 717	*94 409*	427 717	*94 409*	427 717	*94 409*	427 717	*94 409 (22%)*
SRF1	nSeq	-		-		632 798	*4 646 (0.7%)*	-	
	nUnique	-		-		67 497	*922 (1.3%)*	-	
Preclustering	nSeq	987 494	*98 017*	987 494	*98 017*	632 798	*4 646 (0.7%)*	987 494	*98 017*
	nUnique	46 712	*94 409*	46 712	*94 409*	**421**	*320*	46 712	*94 409*
	Time	15min		15min		<1min		15min	
SRF2	nSeq	-		-		-		950 379	*51 734 (5.4%)*
	nUnique	-		-		-		9 597	*9 479*
Chimeras detection	nSeq	930 283	***31 637 (3.4%)***	930 283	*31 637*	632 096	***3 944 (0.6%)***	920 241	***21 596 (2.3%)***
	nUnique	12 335	*12 216*	12 335	*12 216*	169	*68*	2 292	*2 174*
	Time	10min		10min		<1min		2min	
SUB	nSeq	-		300 000	***10 248 (3.4%)***	-		-	
	nUnique	-		4 879	*4 762*	-		-	
Clustering	nSeq	930 283		300 000		632 096		920 241	
	nOTU	6 280	*6 180*	2 659	*2 559*	126	*26*	781	*681*
	Time	70min		6min		<1min		3min	
RareFilter n=20	nSeq	920 136		200 000		632 029		917 386	
	SeqLoss	6.8%		79.7%		36.0%		7.1%	
	nOTU	144	***44***	103	***3***	100	***0***	138	***38***

Simulated chimeric sequences were labelled, thus enabling us to confirm that all the surplus OTUs were chimeric. Although UCHIME detected approximatively 70-80% of the actual chimeras, 3.4% chimeras remained in the total dataset processed by SOP and SUB methods, representing a large number of chimeras (more than 12,000) in the unique sequences. SRF1 drastically reduced the chimera content (to <0.7%) starting from the early filtering step. Chimera detection with UCHIME brought the chimera content down to 0.6% and the rare filter removed the 26 remaining chimeric OTUs, thus procuring chimera-free

datasets. SRF2 was less efficient than SRF1, with 2.3% chimeras after the UCHIME step. In all scenarios, the final rare filter succeeded in removing most of the remaining chimeras but, as mentioned above, this removal was only complete for SRF1, whereas other methods still resulted in surplus OTUs.

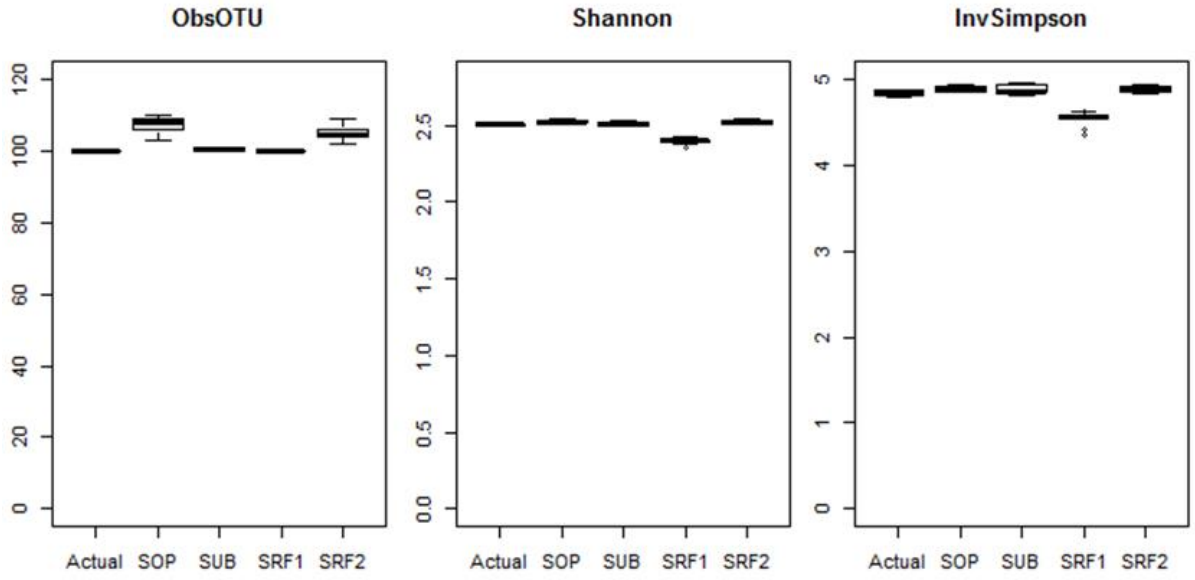


Figure 1: Expected and calculated α-diversity indices for the simulated datasets. Boxplots represent the distribution of results obtained from ten simulated datasets processed with the four methods, SOP, SUB, SRF1 and SRF2.

The α-diversity indices varied depending on the processing method that was applied (**Figure 1**). Shannon and Simpson indices were slightly overestimated with all methods except SRF1, which significantly underestimated both indices. Nevertheless, the OTU distribution, relative abundances and ranks were well conserved (**Figure 2**). For all data processing methods, linear regression of expected and estimated OTU abundances resulted in slopes and correlation coefficients very close to 1. SRF1 was the most divergent method with a slight overestimation of the highly abundant OTUs and a marked underestimation of the less abundant ones. SOP and SRF2 methods displayed the best correlations. The biggest difference observed between the expected and estimated OTU abundances was also produced by SRF1, which overestimates one 42% abundant OTU by 2.4%, resulting in a 5.6% error.
However, Bray-Curtis distances procured good separation of the different datasets, regardless of the processing methods (**Figure 3**). SRF1 was the most divergent method, but its divergence was rather small compared to differences between datasets. Weighted-Unifrac distances and ordinations also lead to similar results for all methods (**Supplementary Data 1**).

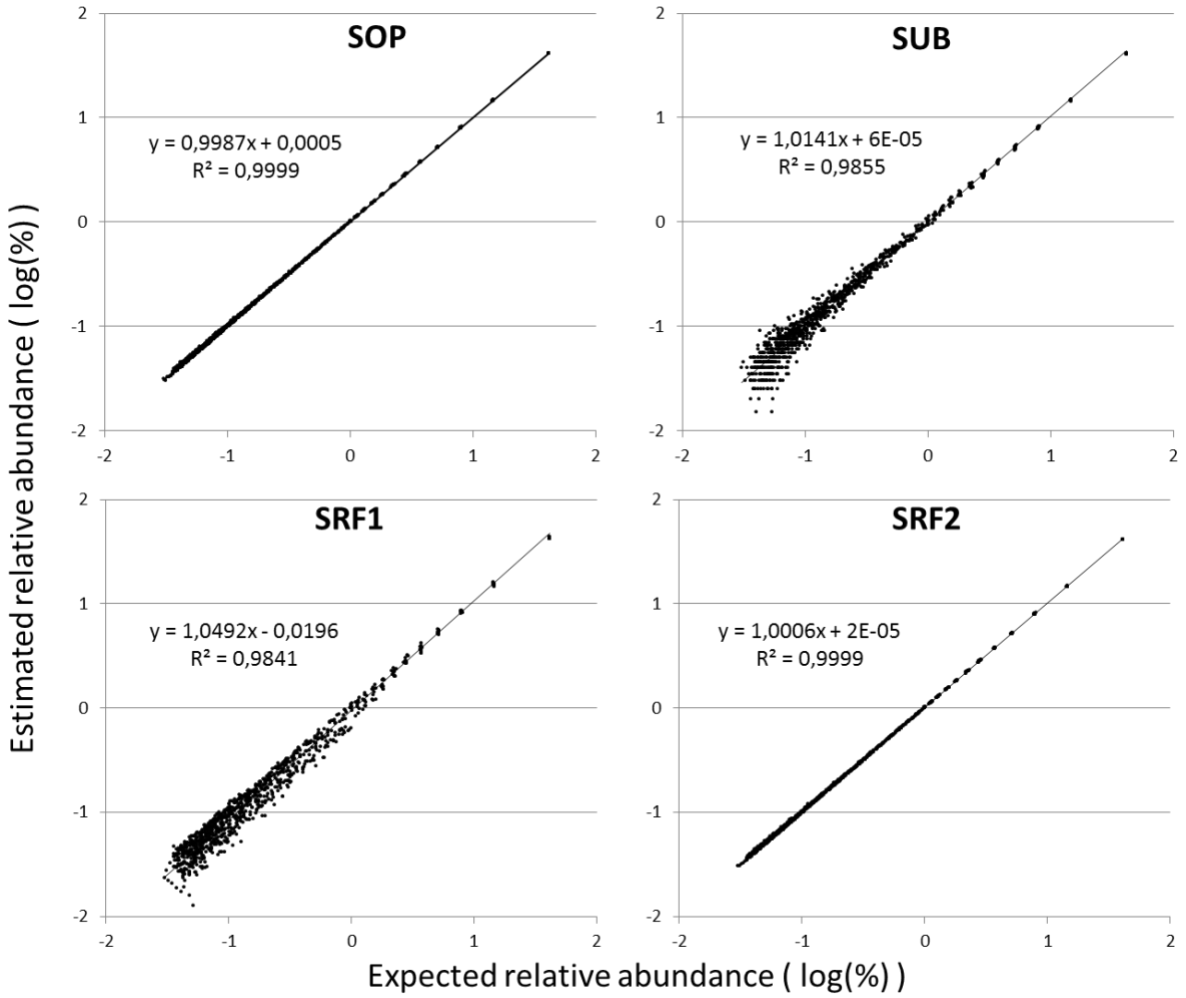


Figure 2: Expected theoretical relative abundance versus estimated relative abundances obtained with SIM data processed by standard SOP, with subsampling SUB, and with singleton read filtering SRF1 and SRF2. Data are presented on a log plot, thus providing better visualization of low abundance OTUs. Continuous lines correspond to expected values.

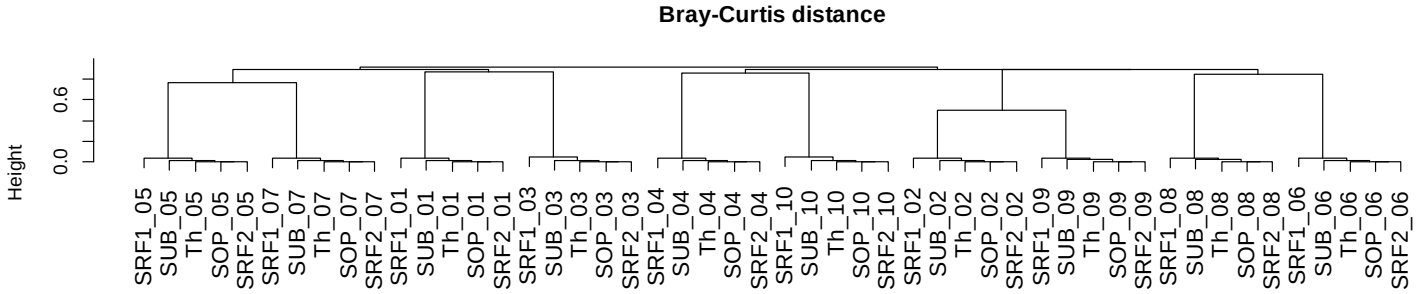


Figure 3: Bray-Curtis distances between theoretical distributions (Th) and processed datasets with SOP, SUB, SRF1 and SRF2 methods, for the ten datasets. Clustering was performed using ward.D2 algorithm of R hclust function.

II.3.2 Mock communities

The V3-V4 16S rRNA gene region of 20-species mock communities HM-782D (even) and HM-783D (staggered) was sequenced by MiSeq Illumina. Mock community data were processed the same way as SIM datasets, except that it was not possible to evaluate the chimera content before the end of the processing. With SOP, which was considered as the reference method, the number of unique sequences was reduced by 83% at the end of the processing whereas SRF1, SRF2 and SUB lead to reductions of respectively, 99.8%, 99.4% and 90% (**Supplementary Data 2A**). As shown with the simulated datasets, such reduction on the number of unique sequences had a high impact on calculation time: SRF1 clustering was performed in only 8s, producing 23 final OTUs; the same number of OTUs was obtained with SUB but in 51min whereas SRF2 and SOP produced, respectively, 28 and 29 OTUs, in 16s and 3h12min, respectively. Alignment of the final OTU sequences with the expected ones allowed to identify the presence of undetected chimeras in datasets processed with SOP and SRF2; it was not the case for SUB and SRF1 methods, which resulted in chimera-free data. However, regardless of the processing methods, 5 to 7 contaminant OTUs were detected in the final data. Such contaminant OTUs corresponded to the most abundant OTUs present in the entire MiSeq run, representing between 0.50% and 0.65% of the final sequences obtained by the different processing methods. Such type of contamination has already been described by (Nelson et al., 2014)). Contaminant OTUs were thus excluded for further downstream analysis. Once contaminants removed, only 17 from the 20 expected OTUs were detected. One undetected OTU is explained by the presence in the mock community of two *Staphylococcus* species with identical V3-V4 16S rRNA regions, they were thus merged in a single OTU. The inspection of the raw sequencing data showed that only few sequences corresponded to *Deinococcus radiodurans* and *Propionibacterium acnes*. *D. radiodurans* present one mismatch with the reverse PCR primer, and was identified only once across the 242,584 raw sequences. *P. acnes* had no mismatch with the primers, but it was only detected 14 times. Due to these low frequencies, probably resulting from unidentified PCR and sequencing bias, these two OTUs were discarded by the rare filter. Hence, based on raw sequencing data, we expect to find 17 OTUs in the mock community, irrespective of the processing method. For further analysis, 17 was thus considered as the actual OTU number of the mock community.

With all methods, α-diversity was lower than the expected values, but was consistent across all methods (**Supplementary Data 2B**). Correlation coefficients between relative abundances

of the expected OTUs (after discarding once chimeric and contaminant OTUs) obtained with SOP and the other methods were strong so the processing methods had no effect on the calculated relative abundances (**Figure 4**). PCoA ordination with Bray-Curtis distances (**Figure 5**) showed that with all methods, communities differ from the expected mock distribution, probably due to sequencing bias. Nevertheless, samples were clustered first by community distribution (even or staggered) and then by technical replicate, irrespectively of the processing method. Hierarchical clustering using Bray-Curtis distances confirmed that differences induced by the processing methods were smaller than the intrinsic variability of technical replicates (**Supplementary Data 2C**).

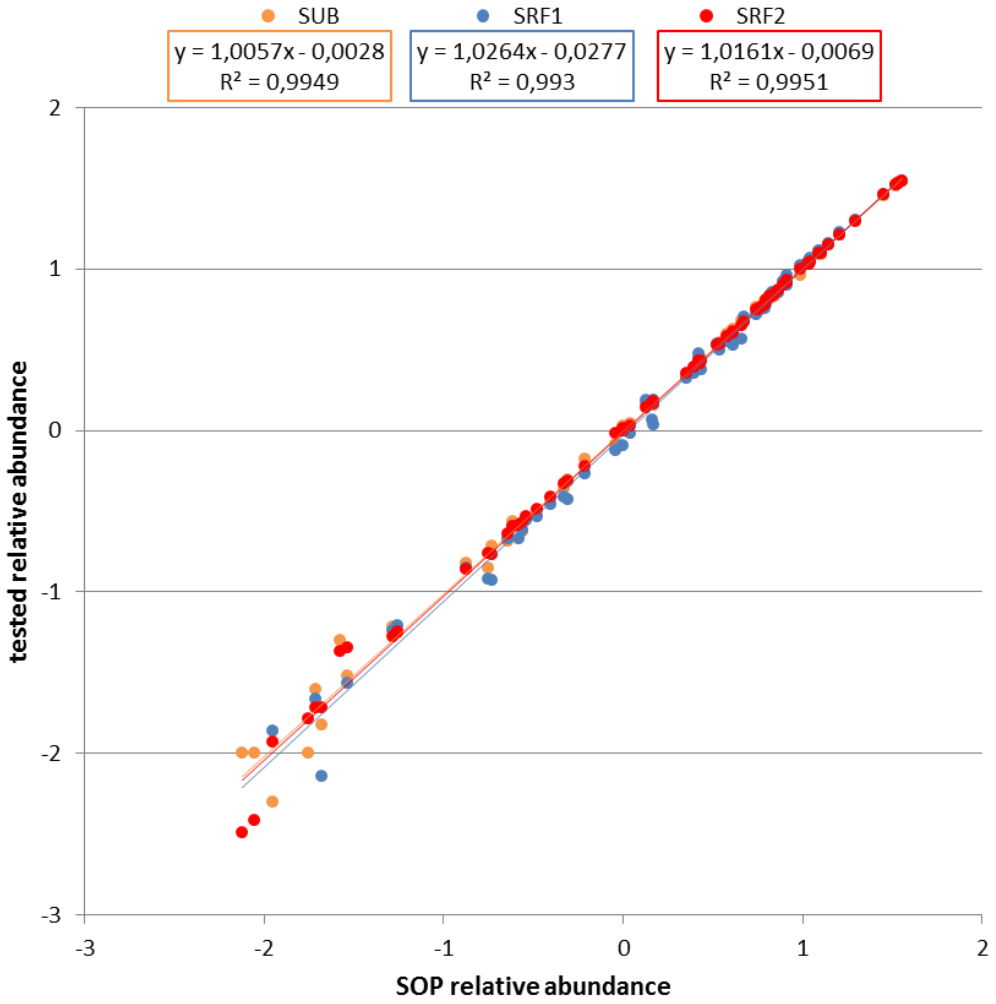


Figure 4: Relative abundance of OTUs in mock communities determined by the SOP method. Unexpected OTUs (chimeras and contaminants) were excluded. Data are presented on a log plot, thus providing better visualization of low abundance OTUs.

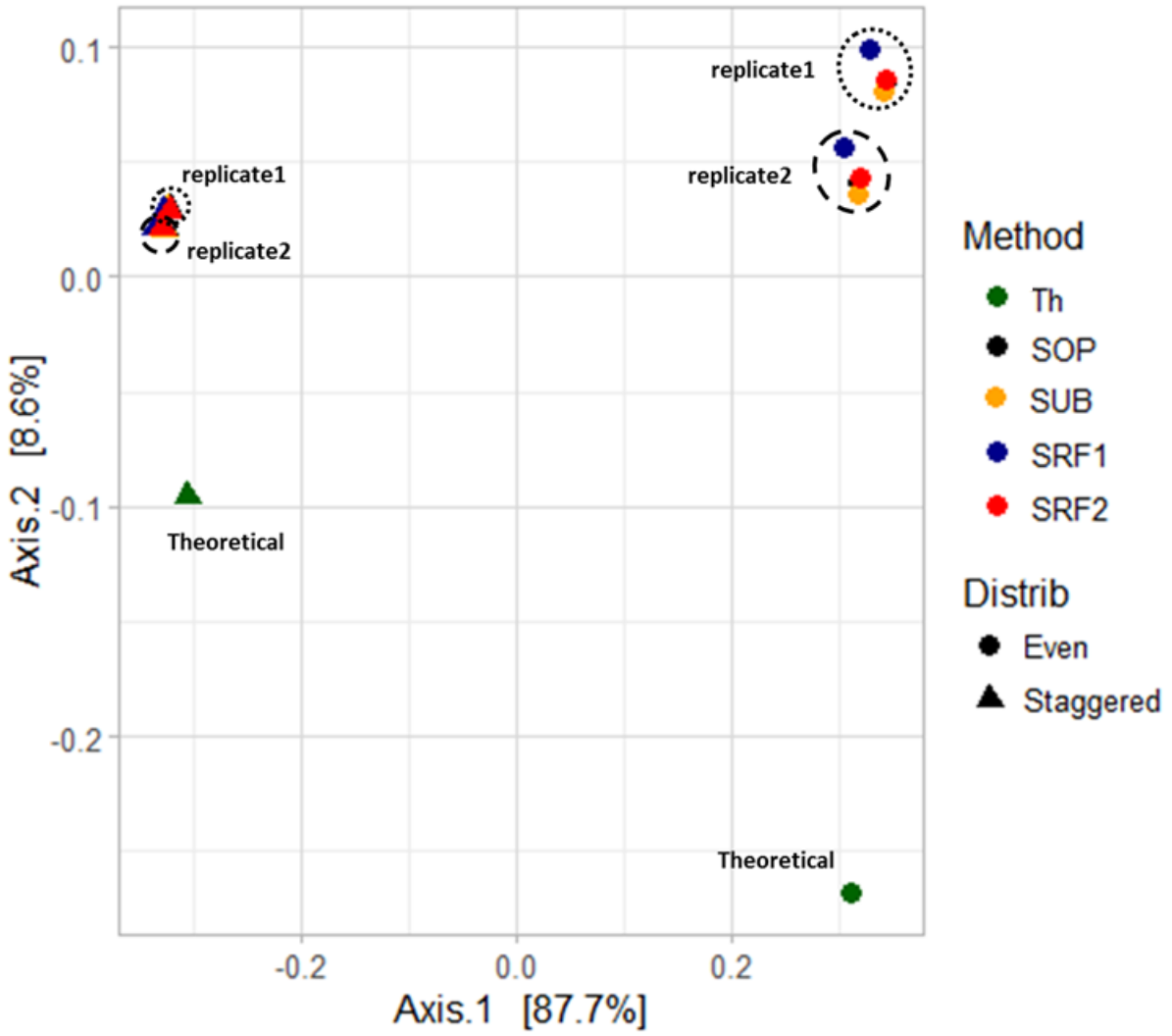


Figure 5: Principle coordinate analysis of OTUs in mock communities based on Bray-Curtis distances, after removal of contaminant OTUs. Dotted and dashed lines encircle replicates 1 and 2 respectively.

II.3.3 Real datasets

To assess the performance of the processing methods with datasets from samples displaying high diversity, four termite gut microflora were sequenced, each one in two biological replicates (pooling 20 guts each). Data from these eight samples was then processed with the four methods. Half of the sequences (53%) were singletons, and were thus filtered with SRF1. Clustering time varied from 8 min (SRF1) to 58 h (SOP), resulting in 637 to 1,114 OTUs (**Supplementary Data 3A**). For some samples, Shannon's and Simpson's diversity index estimated for SRF1 results were slightly lower compared to the other methods (**Supplementary Data 3B**), but PCoA Bray-Curtis ordinations showed that the processing method had no influence on sample separation (**Figure 6**). By integrating the phylogenetic distances, weighted-Unifrac distances improved sample separation and demonstrated that

samples were regrouped first by biological replicate, irrespective of the processing method, and then by the inoculum source (**Supplementary Data 3C**). Weighted-Unifrac distances were thus smaller between processing methods than between biological replicates, meaning that the potential bias due to singleton read filtering was smaller than the differences actually found in biological replicates.

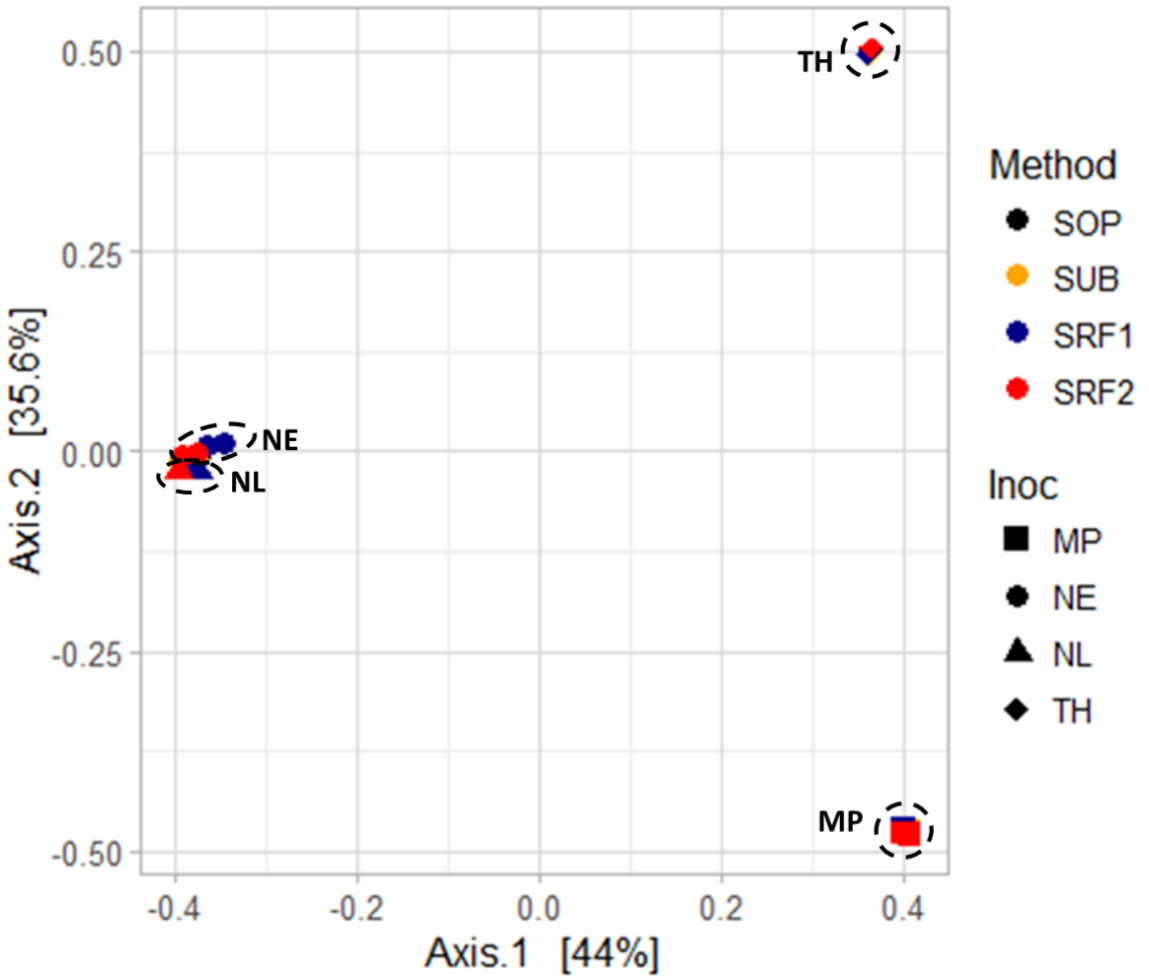


Figure 6: Principle coordinate analysis of OTUs in real communities based on Bray-Curtis distances. Colors correspond to processing methods; shapes denote the sample termite gut origin MP, NE, NL and TH.

II.3.4 Stress datasets

To test the capacity of the different methods to deal with very large numbers of sequences, large datasets of 3.7M (Stress1) and 19M sequences (Stress2) corresponding to two merged MiSeq runs were loaded into the pipelines. Due to large memory requirements that could not be met with some methods, Stress1 was only analyzed with SRF1, SRF2 and SUB methods. The results showed that with 28% sequences filtered, SRF1 reduced the

calculation time more than 500-fold, while producing the same number of OTUs as SUB (**Supplementary Data 4)**. SRF2 was slightly slower than SRF1 and produced 34% more OTUs, but removed only 10% of the total sequences. With real datasets it was not possible to check the final results for remaining chimeras, but the number of filtered total sequences was lower than with data from SIM and mock communities. In fact, comparison of all datasets revealed that the numbers of total and unique sequences apparently did not increase at the same rate, thus the larger the dataset, the smaller the proportion of unique sequences.

Stress2 was only processed by SRF1 because the other methods required several hundred Gbytes of memory for the Mothur average clustering step. Thus, it was not possible to compare SRF1 results between methods, but the change in the number of sequences is similar to the one observed with smaller datasets, with 23% sequence reduction due to SRF. The 18M sequences were processed in 6 days, producing a reasonable number of OTUs (approx. 800).

II.4. Discussion

II.4.1 Chimeras

Simulated datasets enabled us to measure the precise effect of different singleton-filtering methods. In the case of simulated data with 100 species and a random chimera distribution (SIM), it appeared that early filtering of singleton reads (SRF1) procured a 35% read removal and a concomitant decrease of chimera content, from 10% (initial value) to 0.7%. When the singleton-filter (SRF2) was applied at a later stage 2.3% reduction was achieved in combination with chimera detection, whereas with no filtering (SOP and SUB) 3.4% chimeras were undetected. Without exception, all sequences found in the exceeding OTUs were flagged as chimeric, so the “rare biosphere” described in some studies plausibly results from sequencing artefacts, as suggested by previous studies (Huse et al., 2010; Kunin et al., 2010). Furthermore, some chimeric OTUs were abundant enough to pass through the rare filter, except in the case of SRF1 that identified all of 100 expected OTUs; the abundance of rare OTUs was indeed lower than 0.001%. With SRF2 and other methods, the most abundant undetected chimeric OTUs reached 0.018%, which is over the abundance threshold. Overall, filtering the singleton reads appears to be a very good way to eliminate chimeras in the final datasets.

Importantly, chimeras in the simulated data probably do not exactly reproduce the behavior of real chimeras, which could be less randomly distributed. Compared to the random distribution of the chimeras used in this work, real chimeras appear to be focused in a few hotspots and are formed between closer parent sequences (Shin et al., 2014). In this case, rather than singleton chimeras, larger populations of some chimeric sequences may be observed. Nevertheless, the use of mock communities ensured that the amounts of filtered sequences was similar to SIM data and that the final content of chimeric OTUs was close than observed with SIM data. Therefore, it is reasonable to suppose that singleton read filtering (SRF1) applied to data from mock communities mainly removes large populations of chimeric sequences. Moreover, after chimera detection, the number of unique sequences was divided by 5, suggesting that once all the singleton chimeras were removed by SRF1, the detection of the abundant ones was improved. Although some chimeric OTUs remained after SRF1processing, these were rather rare (less than 0.001% abundance) and frequently seen in only one sample. Moreover, these chimeric OTUs were filtered out at the final step, procuring a final chimera-free dataset. With other methods, some low abundance chimeric OTUs remained in the final dataset, passing through the abundance filter. Nevertheless, the number of chimeric OTUs was very low compared to those of non-filtered data, an observation that argues in favor of the use of abundance threshold on OTU tables as suggested in recent publications (Bokulich et al., 2013). Such filters should be seen as a sequencing detection threshold, under which it is not possible to ensure with certainty if an OTU is real or chimeric.

II.4.2 Contaminants

The analysis of *in silico* simulated datasets provided the means to verify that all surplus OTUs were chimeras. In contrast, in mock communities the chimeras were not the only source of spurious OTUs, because contaminant-OTUs were also detected. The presence of such contaminant sequences, coming from other samples in the MiSeq run, has been previously reported in the literature (Nelson et al., 2014). Contaminant OTUs are difficult to detect, because they are real sequences and cannot be identified as artefacts, except when they are detected in *a posteriori* analysis of the sequences arising from a whole MiSeq run. In the case of mock communities, it is possible to identify contaminant OTUs and identify their possible origin by checking the most abundant sequences in the whole MiSeq run. MiSeq sequencing of the mock communities revealed one unexpected OTUs, which was extremely dominant in the whole library and represented 0.33% to 0.4% of the final sequences,

depending of the data processing method. Other smaller contaminants were also abundant in the whole library, and some of them were abundant enough to pass through the rare sequence filters. With SRF1, 6 contaminant OTUs were present, leading to a 25% increase in the expected number of OTUs. The cause of this phenomenon appears to be PCR-independent, because a control sample that was independently amplified, indexed and added into the library just before sequencing was also contaminated. A detection threshold can partially solve the problem by removing the rare contaminants when processing small subsets of the MiSeq run. However, the problem could be tricky to solve if contaminant and contaminated samples are processed simultaneously, because the rare filter is applied over all samples and not per sample. If an OTU is not clearly identified as a contaminant and is abundant in some samples, it becomes difficult to distinguish between a very rare OTU and a contaminant. The only way to have access to the contamination levels is to use pure synthetic DNA as an internal control where any unexpected OTU would be necessarily a contaminant.

II.4.3 Effect on community reconstruction

Even if SRF greatly improved chimera removal and drastically reduced clustering time, it would not be useful if microbial community structure was altered by data processing. Here, Bray-Curtis and weighted-Unifrac analysis showed that all the studied methods described the same relationships between samples, with very small difference in OTU abundances. SRF1 was the fastest method and also the most divergent, but its divergence was small compared to actual differences between samples, irrespective of whether they were biological replicates or technical replicates. Despite an important effect of SRF1 on Shannon's and Simpson's index observed with SIM dataset, it did not impact the relationships between samples. The analysis of data from mock communities revealed that differences induced by the processing methods were smaller than differences between technical replicates. With real insect gut microbial communities, these differences were also smaller than those existing between biological replicates. Therefore, employing either of the four methods described herein had no significant impact on community structure, while singleton read filtering greatly reduced the amount of artefactual sequences in the final datasets. Furthermore, singleton read filtering methods strongly reduces calculation time and improves scalability to huge datasets, as demonstrated with Stress datasets.

II.4.4 Rare filter

This work demonstrates the importance of using a sequencing detection threshold, also called "rare filter". As proposed by(Pylro et al., 2014)), a rare filter appears as the only way to remove undetected chimeras and contaminant sequences arising from other samples of multiplexed runs, both of affect diversity indices. (Pylro et al., 2014)However, such rare filters need to be correctly designed and calibrated. This can be achieved by including a mock community in each sequencing run and estimating the lower filter value that enables the correct recovery of expected sequences from the mock community. When this is not possible, one can also establish either a minimum of sequence abundance in the entire run, or a minimum relative abundance value. Each solution has its advantages and disadvantages. Fixing a filter at 0.005% of total sequence abundance was recently proposed (Bokulich et al., 2013). However, considering a whole MiSeq run of 20M sequences, this 0.005% value represents 1,000 sequences, meaning that every OTU that contained less than 1,000 sequences will be discarded. It is obvious that depending on the number and specific diversity of multiplexed samples, such filter thresholds could induce the removal of real, low abundant OTUs. An alternative way to set a value that is independent of dataset size is based on the minimal sequence abundance in at least one sample. In the case of dissimilar samples with high diversity and only a few OTUs in common, a fixed low threshold avoids filtering real rare OTUs. In the case of increased sequencing depth, the disadvantage of this approach is that the abundance of spurious OTUs can rapidly increase and result in a number of undetected chimeric OTUs that surpasses the threshold. Thus, calibration and use of a rare filter needs to be carefully planned taking into consideration the scientific question and the diversity of the studied samples.

II.5. Material and methods

II.5.1 Simulated data

Simulated sequencing reads (SIM data) were generated using Grinder v 0.5.3 (Angly et al., 2012) with LTP SSU database version 115 (Yarza et al., 2008). Sequences were kept only if they were identified as type species, unambiguous (no N) and matched perfectly to V3-V4 forward (ACGGRAGGCAGCAG) and reverse (TACCAGGGTATCTAATCCT) primers. The 100 sequences that maximized the phylogenetic diversity (based on the neighbor-joining tree distributed with LTP database) were used as reference (see Supplementary data for the species list). Grinder's error parameters were set as follow: error

rate increases linearly from 0.301% to 0.303% per base along the sequences, with 98.6% SNPs and 1.4% indels. These parameters were calibrated by mapping V3-V4 Illumina MiSeq reads from a single strain of *Clostridium perfringens* to its known reference sequence; they were consistent with previously reported values (Schirmer et al., 2015). Chimera proportion was set at 10%, with the Grinder default distribution of 89% bimeras, 11% trimeras and 0.3% quadrimeras, considered as average values (Quince et al., 2011). Chimera breakpoints were uniformly distributed along the sequences. Species abundances of simulated data follow a power law distribution because it best describes abundance distribution in prokaryotes (Gans et al., 2005). Ten samples datasets of 100,000 reads each were generated with a random abundance order and with a maximum/minimum species abundance ratio of 1,000. Cutadapt v1.7.1 (Martin, 2011) was used to trim the primers from the generated reads so the final data look exactly like demultiplexed, contigued and trimmed MiSeq V3-V4 reads. Reads with errors in the primers were discarded.

II.5.2 Sequencing

Genomic DNA from mock communities was obtained through BEI Resources, NIAID, NIH as part of the Human Microbiome Project. The Microbial Mock Community B, Even (HM-782D v5.1L), and Staggered (HM-783D v5.2L) were used directly as DNA template for 16S rRNA sequencing. For each mock distribution, two technical replicates were prepared independently. These mock communities are similar to those used to validate ChimeraSlayer and UPARSE (Haas et al., 2011; Edgar, 2013) except for *Methanobrevibacter* whose amplification failed in these studies and was not included.

Data for the real communities was obtained by sequencing termite gut microbial communities. Termite guts were obtained from the Institut de Recherche et Developpement, (IRD Bondy, France). Two biological replicates of 20 guts each were dissected from *Microcerotermes parvus* (MP), *Nasutitermes ephratae* (NE), *Nasutitermes lujae* (NL) and *Termes hospes* (TH) termite species. Data for the Stress1 and Stress2 datasets were obtained by sequencing samples of real microbial communities issued from lab lignocellulolytic bioreactors inoculated with various inoculum sources. From these samples, DNA was extracted as described previously (Lazuka et al., 2015). Briefly, nucleic acids were co-extracted using a PowerMicrobiome RNA Isolation kit (Mobio Laboratories Inc. Carlsbad) according to the manufacturer's instructions. DNA was then separated and purified using an

AllPrep DNA/RNA MiniKit (Qiagen) following the manufacturer's instruction. DNA quality was checked by electrophoresis on a 1% agarose gel and ethidium bromide staining.

DNA from mock and real microbial communities was quantified by NanoDrop 1000 spectrophotometer (ThermoScientific) and adjusted to 2ng/µL. 16S rRNA was amplified and sequenced using MiSeq Illumina sequencing performed by the GenoToul Genomics and Transcriptomics facility (GeT-PlaGe, Auzeville, France). A 460 bp fragment corresponding to the V3-V4 16S rRNA gene region was amplified from genomic DNA samples using bacterial universal primers F343 and R784. Primer sequences and PCR conditions are detailed in in Supplementary Information. The purified PCR products were equimolarly pooled and loaded onto the Illumina MiSeq cartridge using v3 reagents according to the manufacturer's instructions.

II.5.3 Data processing

Sequencing Illumina MiSeq data delivered by GeT-PlaGe were already demultiplexed and contigued with Flash v1.2.6 (Magoč and Salzberg, 2011) with a minimum of 110bp overlap and less than 0.1 mismatches. A homemade script was used to convert the fastq file to a fasta file and create the corresponding Mothur group file. Four processing methods were tested: three of them were variants of the main pipeline. This main pipeline roughly consisted in Mothur Stantard Operating Procedure (Kozich et al., 2013) for Illumina data, so it is hereafter called SOP. Mothur was used to trim the primers (except for simulated data) with no mismatch allowed, and to remove the N-containing sequences or those displaying erroneous length (>380 and <460bp). Sequences were aligned versus the expected V3-V4 region using a multi-alignment database approach; only the sequences aligning to the expected region were conserved. Unique sequences were de-noised using Mothur's preclustering tool with a diffs parameter set to 5, meaning that all sequences with a maximum of 5 nt differences with an abundant sequence were merged with the abundant one. Unique sequences which pass through Mothur UCHIME self-reference implementation were clustered using Mothur's clustering average-distance tool. Rare OTUs containing less than 20 sequences (equivalent to less than 0.005%) were then filtered out. OTUs were taxonomically assigned based on the consensus sequence of each OTU, using Mothur's implementation of RDP Classifier and LTP SSU database version 115 (Yarza et al., 2008). In order to reduce the number of sequences to process, three data processing methods were tested: the first one included a random subsampling step (SUB); the second one consisted in a singleton read filter applied before

preclustering (SRF1) and, the last one applied the singleton filter after preclustering (SRF2). All processing methods were performed with Mothur functions. OTU tables were constructed with Mothur, and OTU fasta files were generated using the most abundant sequence of each OTU. Phylogenetic trees were generated using ClustalO (Sievers et al., 2014) and raxmlHPC (Stamatakis, 2014).

II.5.4 Diversity analyses

OTU tables, taxonomy files and phylogenetic trees were imported into R v3.1.2 using the Phyloseq package version 1.10.0 (McMurdie and Holmes, 2013). Diversity indexes (Shannon and Simpson) and distances between samples (Bray-Curtis or weigthed-Unifrac) were all calculated using Phyloseq. Distances were clustered with hclust function of the *stats* package v3.1.2. Heatmaps were drawn using the pheatmap v1.0.7 R package.

II.6. Acknowledgments

This research was supported by the French National Institure for Agronomical Reaseach (INRA) and the Region Languedoc-Roussillon Midi-Pyrénées. The authors thank the Genomics and Transcriptomics (GeT) platform for their help with sequencing.

II.7. References

Angly, F.E., Willner, D., Rohwer, F., Hugenholtz, P., and Tyson, G.W. (2012). Grinder: a versatile amplicon and shotgun sequence simulator. Nucleic Acids Res. gks251.

Bokulich, N.A., Subramanian, S., Faith, J.J., Gevers, D., Gordon, J.I., Knight, R., Mills, D.A., and Caporaso, J.G. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat. Methods *10*, 57–59.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., Fierer, N., and Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc. Natl. Acad. Sci. U. S. A. *108*, 4516–4522.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461.

Edgar, R.C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat. Methods *10*, 996–998.

Gans, J., Wolinsky, M., and Dunbar, J. (2005). Computational Improvements Reveal Great Bacterial Diversity and High Metal Toxicity in Soil. Science *309*, 1387–1390.

González, I., Ayuso-Sacido, A., Anderson, A., and Genilloud, O. (2005). Actinomycetes isolated from lichens: Evaluation of their diversity and detection of biosynthetic gene sequences. FEMS Microbiol. Ecol. *54*, 401–415.

Haas, B.J., Gevers, D., Earl, A.M., Feldgarden, M., Ward, D.V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S.K., Sodergren, E., et al. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res. *21*, 494–504.

Hugenholtz, P., Goebel, B.M., and Pace, N.R. (1998). Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity. J. Bacteriol. *180*, 4765–4774.

Huse, S.M., Welch, D.M., Morrison, H.G., and Sogin, M.L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ. Microbiol. *12*, 1889–1898.

Jumpponen, A., and Johnson, L.C. (2005). Can rDNA analyses of diverse fungal communities in soil and roots detect effects of environmental manipulations—a case study from tallgrass prairie. Mycologia *97*, 1177–1194.

Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., and Schloss, P.D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl. Environ. Microbiol. *79*, 5112–5120.

Kunin, V., Engelbrektson, A., Ochman, H., and Hugenholtz, P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ. Microbiol. *12*, 118–123.

Magoč, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinforma. Oxf. Engl. *27*, 2957–2963.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12.

McMurdie, P.J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PloS One *8*, e61217.

McMurdie, P.J., and Holmes, S. (2014). Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Comput Biol *10*, e1003531.

Medinger, R., Nolte, V., Pandey, R.V., Jost, S., Ottenwälder, B., Schlötterer, C., and Boenigk, J. (2010). Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. Mol. Ecol. *19*, 32–40.

Meeus, I., Parmentier, L., Billiet, A., Maebe, K., Van Nieuwerburgh, F., Deforce, D., Wäckers, F., Vandamme, P., and Smagghe, G. (2015). 16S rRNA Amplicon Sequencing Demonstrates that Indoor-Reared Bumblebees (Bombus terrestris) Harbor a Core Subset of Bacteria Normally Associated with the Wild Host. PLoS ONE *10*.

Nelson, M.C., Morrison, H.G., Benjamino, J., Grim, S.L., and Graf, J. (2014). Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. PloS One *9*, e94249.

Pylro, V.S., Roesch, L.F.W., Morais, D.K., Clark, I.M., Hirsch, P.R., and Tótola, M.R. (2014). Data analysis for 16S microbial profiling from different benchtop sequencing platforms. J. Microbiol. Methods *107*, 30–37.

Quince, C., Lanzen, A., Davenport, R.J., and Turnbaugh, P.J. (2011). Removing noise from pyrosequenced amplicons. BMC Bioinformatics *12*, 38.

Schirmer, M., Ijaz, U.Z., D'Amore, R., Hall, N., Sloan, W.T., and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res.

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., et al. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl. Environ. Microbiol. *75*, 7537–7541.

Shin, S., Lee, T.K., Han, J.M., and Park, J. (2014). Regional effects on chimera formation in 454 pyrosequenced amplicons from a mock community. J. Microbiol. Seoul Korea *52*, 566–573.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., et al. (2014). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539–539.

Stamatakis, A. (2014). RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics btu033.

Tedersoo, L., Nilsson, R.H., Abarenkov, K., Jairus, T., Sadam, A., Saar, I., Bahram, M., Bechem, E., Chuyong, G., and Kõljalg, U. (2010). 454 Pyrosequencing and Sanger

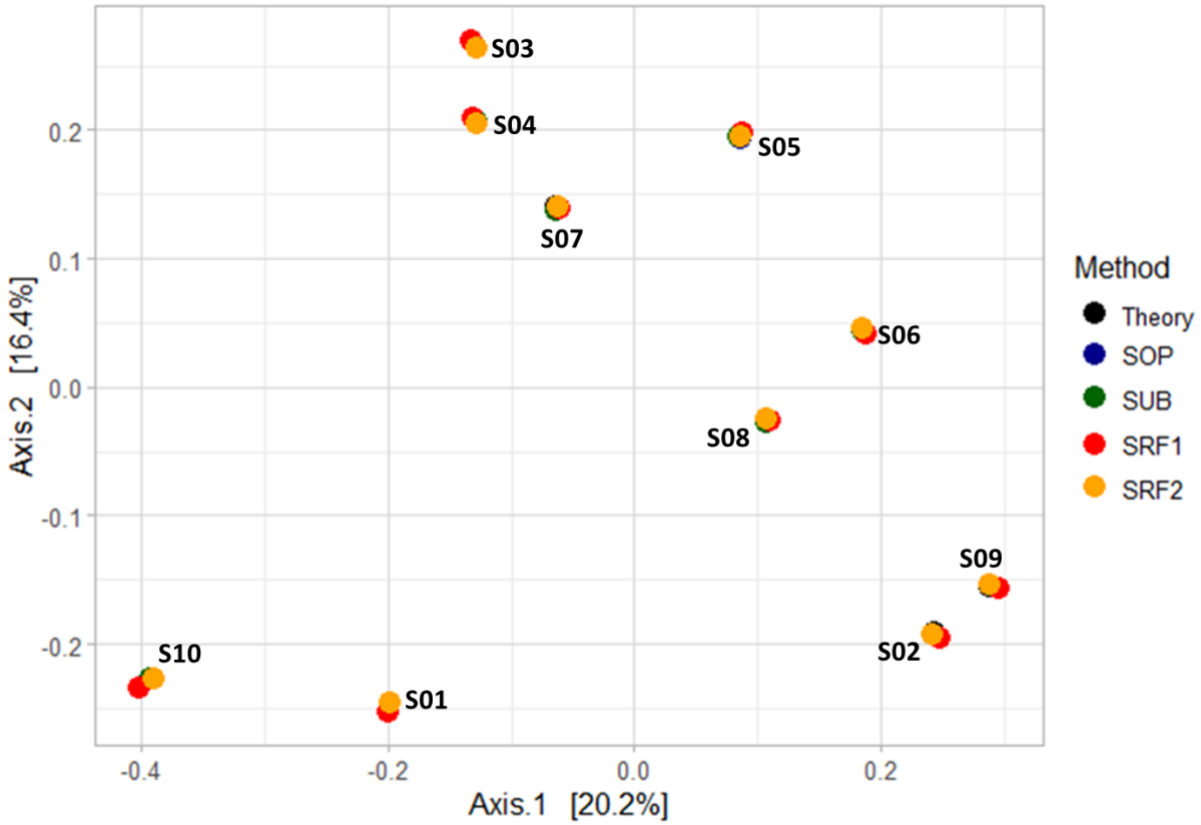
sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. New Phytol. *188*, 291–301.

Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K.-H., Ludwig, W., Glöckner, F.O., and Rosselló-Móra, R. (2008). The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst. Appl. Microbiol. *31*, 241–250.

II.8. Supplementary data

Supplementary Data 1: Simulated data

Weighted-Unifrac distances between samples were calculated and a PCoA ordination was performed using R Phyloseq package. Each color represents a processing method, including theoretical distribution (black). Points are grouped by dataset, irrespective of the processing method, so the methods have a small impact on communities' relationships.



Supplementary Data 2: mock communities

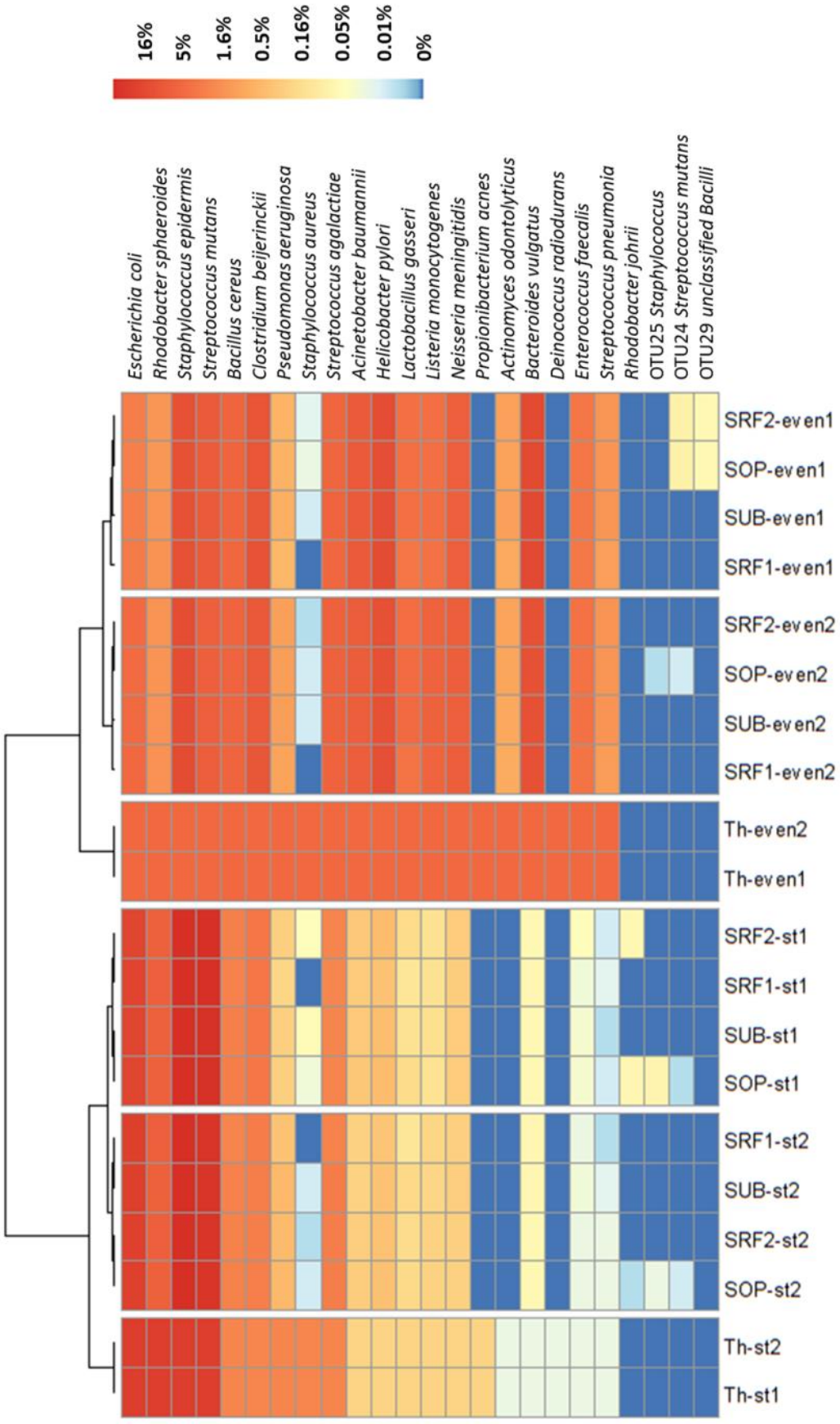
A: Summary of mock community's data processing by the four methods. The total number of sequences, number of unique sequences and sequences lost at each step are detailed. Final chimera and contaminant content is shown in the last row.

Step	**Method**	number of sequences							
		SOP	*Seq loss*	**SUB**	*Seq loss*	**SRF1**	*Seq loss*	**SRF2**	*Seq loss*
Filtering	nSeq	242 584		242 584		242 584		242 584	
	nUniq	121 003	*50%*	121 003	*50%*	121 003	*50%*	121 003	*50%*
SRF1	nSeq	-		-		141 086	*-42%*	-	
	nUniq	-		-		15 003	*-88%*	-	
Preclustering	nSeq	242 584		242 584		141 086	*-42%*	242 584	*0%*
	nUniq	24 666	*-80%*	24 666	*-80%*	**1 003**	*-99,2%*	24 666	*-80%*
	Time	2min		2min		5s		2min	
SRF2	nSeq	-		-		-		220 234	*-9%*
	nUniq	-		-		-		2 316	*-98,1%*
Chimeras detection	nSeq	220 209	***-9%***	220 209	*-9%*	132 603	***-45%***	200 300	***-17%***
	nUniq	20 649	*-83%*	20 649	*-83%*	186	*-99,8%*	740	*-99,4%*
	Time	27min		27min		8s		46s	
SUB	nSeq	-		120 000	***-51%***	-		-	
	nUniq	-		11 559	*-90%*	-		-	
Clustering	nSeq	220 209	*-9%*	120 000	*-51%*	132 603	*-45%*	200 300	*-17%*
	nOTU	4 709		2 763		69		103	
	Time	3h12		51min		8s		16s	
RareFilter n=20	nSeq	214 437	*-12%*	80 000	*-67%*	132 452	*-45%*	200 021	*-18%*
	nOTU	**29** *Actual : 17* **4 cont. OTU** **chimeras**		**23** *Actual : 17* **0 cont. OTU** **chimeras**		**23** *Actual : 17* **0 cont. OTU** **chimeras**		**28** *Actual : 17* **3 cont. OTU** **chimeras**	

B: Shannon's and Simpson's diversity index values for the two mock distributions

	Expected	**SOP**	**SUB**	**SRF1**	**SRF2**
ObsOTUs (ev)	17	29±0	23±0	22.5±0.5	26±1.0
Shannon (ev)	2.996	2.654±0.023	2.652±0.026	2.615±0.031	2.652±0.024
InvSimpson (ev)	20	12.36±0.51	12.40±0.54	11.86±0.55	12.34±0.53
ObsOTUs (st)	17	28±1	22±0	21.5±0.5	24±1.0
Shannon (st)	1.839	1.657±0.003	1.645±0.011	1.627±0.005	1.654±0.003
InvSimpson (st)	5.145	3.970±0.042	3.967±0.066	3.935±0.042	3.98±0.046

C: Bray-Curtis hierarchical clustering of mock samples with even (even1 and even2) or staggered distribution (st1 and st2) and the four processing methods. Colors correspond to relative abundances with each method. Samples were ordered according to their Bray-Curtis distance, clustered with ward.D2 algorithm.



Supplementary Data 3: real samples

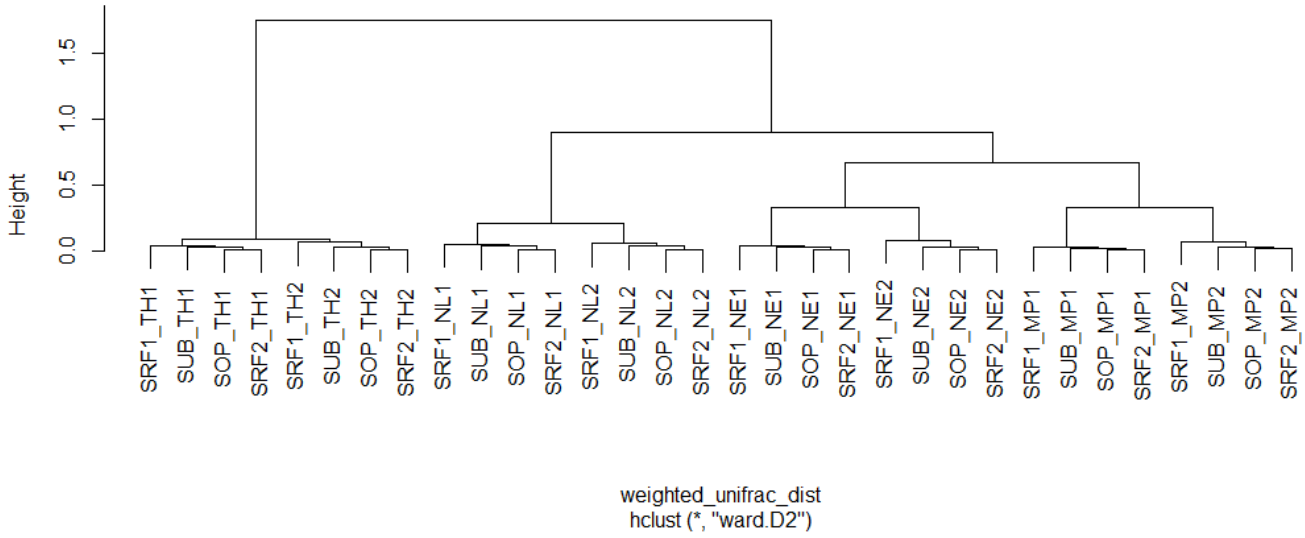
A: Summary of real samples processing by the four methods. The total number of sequences, number of unique sequences, and sequences lost at each step are detailed.

Step	**Method**	number of sequences							
		SOP	*Seq loss*	**SUB**	*Seq loss*	**SRF1**	*Seq loss*	**SRF2**	*Seq loss*
Filtering	nSeq	400 978		400 978		400 978		400 978	
	nUnique	258 522		258 522		258 522		258 522	
SRF1	nSeq	-		-		186 604	*-53%*	-	
	nUnique	-		-		41 517	*-84%*	-	
Preclustering	nSeq	400 987		400 987		186 604	*-42%*	400 987	
	nUnique	83 341	*-68%*	83 341	*-68%*	**12 871**	*-95,0%*	83 341	*-68%*
	Time	6min		6min		12s		6min	
SRF2	nSeq	-		-		-		339 130	*-15%*
	nUnique	-		-		-		21 493	*-91,60%*
Chimeras detection	nSeq	337 468	***-16%***	337 468	***-16%***	163 033	***-59%***	291 202	***-27%***
	nUnique	58 378	*-77%*	58 378	*-77%*	7 433	*-97,10%*	12 115	*-95,30%*
	Time	24min		24min		29s		3min	
SUB	nSeq	-		120 000	***-70%***	-		-	
	nUnique	-		26 069	*-90%*	-		-	
Clustering	nSeq	337 468	*-16%*	120 000	*-70%*	163 033	***-59%***	291 202	*-27%*
	nOTU	**24 090**		**10 988**		3 323		4 385	
	Time	**58h**		6h56		8min		43min	
RareFilter n=20	nSeq	295 147	*-26%*	80 000	*-80%*	149 585	*-63%*	273 163	*-32%*
	nOTU	**1 114**		**692**		**637**		**965**	

B: Shannon's and Simpson's diversity index values for the two mock distributions

		SOP	**SUB**	**SRF1**	**SRF2**
Shannon	MP	3,69±0,04	3,56±0,05	3,25±0,11	3,45±0,06
	NE	3,66±0,03	3,48±0,04	3,41±0,11	3,58±0,03
	NL	3,62±0,05	3,42±0,07	3,31±0,01	3,56±0,04
	TH	5,39±0,01	5,12±0,02	4,95±0,07	5,31±0,02
Simpson	MP	15,91±1,88	15,29±1,92	10,4±2,02	11,52±1,57
	NE	14,12±0,19	13,19±0,02	13,22±1,74	13,64±0,13
	NL	14,13±0,01	13,19±0,21	11,86±0,79	13,9±0,04
	TH	84,57±1,19	72,96±3,76	55,53±4,42	81,28±0,59

C: Weighted-Unifrac distances between samples. SOP, SUB, SRF1 and SRF2 correspond to the processing method applied. MP, NE, NL and TH correspond to sample termite gut origin. Numbers 1 and 2 correspond to the biological replicates 1 and 2.



Supplementary Data 4: stress datasets

Number of sequences along Stress 1 & 2 data processing. Total number of sequences and sequences lost at each step are detailed. Clustering time is detailed in the Clustering row.

Step	**Method**	**Stress1**						**Stress2**	
		SUB	*Seq loss*	**SRF1**	*Seq loss*	**SRF2**	*Seq loss*	**SRF1**	*Seq loss*
Filtering	nSeq	3 432 509		3 432 509		3 432 509		17 894 384	
	nUnique	1 151 370	*34%*	1 151 370	*34%*	1 151 370	*34%*	5 224 590	
SRF1	nSeq	-		2522200	*-26%*	-		13 799 153	*-23%*
	nUnique	-		241061	*-79%*	-		1 129 359	*-78%*
Preclustering	nSeq	3 419 203	*0*	2 515 682	*0*	3 419 203	*0%*	13 759 283	*-23%*
	nUnique	265 289	*-77%*	21 038	*-98%*	265 289	*-77,0%*	141 779	*-97%*
SRF2	nSeq	-		-		3 184 689	*-7%*	-	
	nUnique	-		-		30 775	*-97%*	-	
Chimeras Detection	nSeq	3 319 030	*-3%*	2 459 521	*-28%*	3 100 029	*-10%*	13 147 896	*-27%*
	nUnique	233 401	*-80%*	9 117	*-99%*	14432	*-98,7%*	57 299	*-98,9%*
SUB	nSeq	1650000	*-52%*	-		-		-	
	nUnique	117468	*-90%*	-		-		-	
Clustering	nSeq	1 650 000	***-52%***	2 459 521	***-28%***	3 100 029	***-10%***	13 147 896	***-26,5%***
	nOTU	35 775		1 669		2069		7 389	
	Time	23 days		1 hour		3 hours		6 days	
RareFilter n=20	nSeq	1 100 000	*-68%*	2 452 485	*-29%*	3 093 074	*-10%*	13 095 814	*-26,8%*
	nOTU	**429**		**435**		**587**		**788**	

Supplementary Information: simulated data species

ID	Phylum	Genus	Species
AAOA01000004	Proteobacteria	*Congregibacter*	*litoralis*
AB071324	Aquificae	*Sulfurihydrogenibium*	*subterraneum*
AB089844	Firmicutes	*Sulfobacillus*	*thermosulfidooxidans*
AB106353	Firmicutes	*Tepidanaerobacter*	*syntrophicus*
AB120294	Aquificae	*Hydrogenivirga*	*caldilitoris*
AB110421	Proteobacteria	*Saccharibacter*	*floricola*
AB176554	Proteobacteria	*Fangia*	*hongkongensis*
AB210824	Firmicutes	*Sharpea*	*azabuensis*
AB241105	Actinobacteria	*Bifidobacterium*	*criceti*
AB193261	Actinobacteria	*Patulibacter*	*minatonensis*
AB231858	Nitrospira	*Thermodesulfovibrio*	*yellowstonii*
AB272165	Bacteroidetes	*Persicitalea*	*jodogahamensis*
AB331888	Verrucomicrobia	*Roseibacillus*	*ishigakijimensis*
AB298731	Actinobacteria	*Propioniciclava*	*tarda*
AB428365	Caldiserica	*Caldisericum*	*exile*
AB298736	Bacteroidetes	*Anaerocella*	*delicata*
AB449109	Firmicutes	*Natribacillus*	*halophilus*
AB558581	Lentisphaerae	*Oligosphaera*	*ethanolica*
AB525415	Proteobacteria	*Eikenella*	*corrodens*
AB594446	Actinobacteria	*Branchiibius*	*cervicis*
AB558927	Proteobacteria	*Pacificibacter*	*maritimus*
AE009951	Fusobacteria	*Fusobacterium*	*nucleatum*
AE017126	Cyanobacteria	*Prochlorococcus*	*marinus*
AF073450	Tenericutes	*Ureaplasma*	*urealyticum*
AF084852	Proteobacteria	*Peredibacter*	*starrii*
AF133538	Proteobacteria	*Oligella*	*urethralis*
AF170103	Chlorobi	*Chloroherpeton*	*thalassium*
AF228001	Proteobacteria	*Gallibacterium*	*anatis*
AF146526	Deferribacteres	*Denitrovibrio*	*acetiphilus*
AF357916	Spirochaetes	*Sphaerochaeta*	*globosa*
AF418180	Proteobacteria	*Desulfobacter*	*postgatei*
AF548373	Firmicutes	*Allisonella*	*histaminiformans*
AJ010963	Firmicutes	*Centipeda*	*periodontii*
AF537211	Firmicutes	*Filifactor*	*villosus*
AJ292759	Proteobacteria	*Bdellovibrio*	*bacteriovorus*
AJ247194	Proteobacteria	*Asticcacaulis*	*excentricus*
AJ308318	Proteobacteria	*Azorhizophilus*	*paspali*
AJ417075	Firmicutes	*Allobaculum*	*stercoricanis*
AJ413954	Firmicutes	*Faecalibacterium*	*prausnitzii*
AJ428402	Actinobacteria	*Varibaculum*	*cambriense*
AJ496806	Firmicutes	*Alicyclobacillus*	*acidocaldarius*
AJ458420	Firmicutes	*Clostridium*	*butyricum*
AJ439543	Firmicutes	*Lactovum*	*miscens*

AJ496032	Fibrobacteres	*Fibrobacter*	*succinogenes*
AM162405	Acidobacteria	*Bryobacter*	*aggregatus*
AM490846	Elusimicrobia	*Elusimicrobium*	*minutum*
AM747811	Actinobacteria	*Enterorhabdus*	*mucosicola*
AM749780	Armatimonadetes	*Chthonomonas*	*calidirosea*
AM941746	Proteobacteria	*Kushneria*	*aurantia*
AY455809	Proteobacteria	*Thiomonas*	*intermedia*
AY557615	Proteobacteria	*Silanimonas*	*lenta*
AY773949	Firmicutes	*Lactobacillus*	*delbrueckii*
CP000414	Firmicutes	*Leuconostoc*	*mesenteroides*
CP000027	Chloroflexi	*Dehalococcoides*	*mccartyi*
CP000771	Thermotogae	*Fervidobacterium*	*nodosum*
CP000875	Chloroflexi	*Herpetosiphon*	*aurantiacus*
CP001685	Fusobacteria	*Leptotrichia*	*buccalis*
CP002122	Bacteroidetes	*Prevotella*	*melaninogenica*
CU463952	Synergistetes	*Cloacibacillus*	*evryensis*
D89067	Chlamydiae	*Chlamydia*	*trachomatis*
DQ019166	Firmicutes	*Exiguobacterium*	*aurantiacum*
DQ327663	Proteobacteria	*Verminephrobacter*	*eiseniae*
DQ457019	Bacteroidetes	*Niabella*	*aurantiaca*
DQ768123	Proteobacteria	*Pyxidicoccus*	*fallax*
EF067861	Proteobacteria	*Solimonas*	*soli*
DQ680836	Bacteroidetes	*Parapedobacter*	*koreensis*
EF217419	Bacteroidetes	*Hyunsoonleella*	*jejuensis*
EF660760	Proteobacteria	*Henriciella*	*marina*
EU327343	Firmicutes	*Halanaerobaculum*	*tunisiense*
EU240886	Firmicutes	*Nosocomiicoccus*	*ampullae*
EU564841	Proteobacteria	*Sphingomicrobium*	*lutaoense*
EU660053	Actinobacteria	*Haloglycomyces*	*albus*
FJ482231	Proteobacteria	*Thiohalobacter*	*thiocyanaticus*
FJ796700	Firmicutes	*Lachnoanaerobaculum*	*umeaense*
FN421478	Bacteroidetes	*Fontibacter*	*flavus*
FR753034	Proteobacteria	*Tardiphaga*	*robiniae*
FR733705	Thermotogae	*Petrotoga*	*miotherma*
GQ355622	Bacteroidetes	*Fibrisoma*	*limi*
GQ857549	Proteobacteria	*Diplorickettsia*	*massiliensis*
GQ922842	Chrysiogenetes	*Desulfurispira*	*natronophila*
GU575117	Proteobacteria	*Pseudahrensia*	*aquimaris*
HE614680	Bacteroidetes	*Cruoricaptor*	*ignavus*
JF262044	Firmicutes	*Caloribacterium*	*cisternae*
HQ832501	Actinobacteria	*Gryllotalpicola*	*koreensis*
HQ537484	Proteobacteria	*Aliidiomarina*	*taiwanensis*
JN605361	Proteobacteria	*Limimonas*	*halophila*
L09178	Firmicutes	*Caldicellulosiruptor*	*saccharolyticus*
JN880417	Planctomycetes	*Telmatocola*	*sphagniphila*
JQ080912	Proteobacteria	*Galenea*	*microaerophila*

U01330	Proteobacteria	*Helicobacter*	*pylori*
U25627	Proteobacteria	*Thermodesulforhabdus*	*norvegica*
U39399	Proteobacteria	*Psychrobacter*	*immobilis*
X69335	Firmicutes	*Coprothermobacter*	*proteolyticus*
U93332	Actinobacteria	*Kitasatospora*	*setae*
X58890	Actinobacteria	*Mycobacterium*	*tuberculosis*
Y11332	Deinococcus-Thermus	*Deinococcus*	*radiodurans*
Y17600	Proteobacteria	*Succinivibrio*	*dextrinosolvens*
Z12817	Spirochaetes	*Leptospira*	*interrogans*
Z22781	Spirochaetes	*Brachyspira*	*aalborgi*
Z38007	Actinobacteria	*Saccharomonospora*	*viridis*

PCR conditions

PCR1:

343F= 5'-CTT TCC CTA CAC GAC GCT CTT CCG ATC TAC GGR AGG CAG CAG-3'
784R= 5'-GGA GTT CAG ACG TGT GCT CTT CCG ATC TTA CCA GGG TAT CTA ATC CT-3'

PCR was performed in 50µl reaction mixture containing 1X PCR buffer, 2.5U MTP Taq DNA Polymerase (Sigma), 0.2mM of each dNTP, 0.5mM of each primer and 2ng of extracted DNA. After 30 amplification cycles of 94°C-65°C-70°C, one minute each step, amplicons were purified using magnetic beads and quantified by NanoDrop 1000 spectrophotometer.

PCR2:

FP2= 5'-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC-3'
RP2= 5'-CAA GCA GAA GAC GGC ATA CGA GAT-**index**-GTG ACT GGA GTT CAG ACG TGT-3'

PCR was performed in 50µl reaction mixture containing 1X PCR buffer, 2.5U MTP Taq DNA Polymerase (Sigma), 0.2mM of each dNTP, 0.5mM of each primer and 15ng of previously amplified DNA. After 12 amplification cycles of 94°C-65°C-72°C, one minute each step, amplicons were purified using magnetic beads and quantified by Nanodrop 1000 spectrophotometer.

III. Conclusion du chapitre

La filtration des singletons dès les premières étapes du séquençage, et notamment avant l'étape critique du clustering, apparait comme une méthode permettant d'augmenter considérablement les performances et les capacités des outils classiques comme Mothur. De plus, elle n'altère pas les résultats obtenus, et peut même les améliorer. En effet, les différences induites par les différentes méthodes comparées étant plus petites que les différences entre réplicats techniques et biologiques. Les relations entre échantillons sont donc conservées par la filtration des singletons, qui de plus réduit le nombre de faux taxons produits.

La filtration des singletons permet donc de traiter les jeux de données complets produits par un séquenceur Illumina MiSeq pour l'analyse simultanée des séquences provenant de centaines d'échantillons à comparer. C'est cette méthode qui est utilisée dans les chapitres suivants.

Cependant, même s'ils sont minimes, la méthode introduit des biais, et d'autres solutions sont en cours de développement avec des outils de clustering capables de gérer un très grand nombre de séquence. C'est notamment le cas de FROGS (Find Rapidly OTUs with Galaxy Solution), qui associe à un nouvel outil la facilité d'utilisation d'une interface web Galaxy. Son développement, en collaboration avec l'équipe NED de l'INRA Toulouse fera prochainement l'objet d'une publication en co-auteur. FROGS a déjà fait l'objet de plusieurs présentations, sous forme de poster ou d'oral, à des congrès nationaux ou internationaux (notamment JOBIM Clermont-Ferrand, Juillet 2015, et Pathobiome Paris, Juin 2015).

CHAPITRE IV

STABILISATION ET CARACTERISATION D'UN INOCULUM ISSU DE RUMEN BOVIN

CHAPITRE IV : STABILISATION ET CARACTERISATION D'UN INOCULUM ISSU DE RUMEN BOVIN

I. Introduction

La lignocellulose est le plus grand réservoir de carbone renouvelable au monde, et représente une bonne alternative aux ressources fossiles. Cependant, sa valorisation est encore très limitée du fait de la difficulté à la dégrader. En effet, les méthodes enzymatiques classiques, utilisées pour la production de biocarburants, sont insuffisantes dans le cas de substrats lignocellulosique. D'autres approches, comme la plateforme carboxylate, reposant sur l'utilisation de communautés microbiennes complexes pour faire face à la complexité des substrats utilisés, sont maintenant bien décrites. Si l'utilisation de techniques de prétraitement est indispensable à améliorer la dégradation d'un substrat, l'amélioration des rendements passe également par la sélection de communautés microbiennes plus efficaces.

De nombreux travaux s'attèlent à cette tâche, et des communautés opérant en conditions aérobies ou anaérobies ont été sélectionnées sur des substrats simples, mais aussi complexes (paille de blé, de riz, résidus de maïs ou de fauchage), prétraités ou non (Guo et al., 2010; Reddy et al., 2011; Gao et al., 2013). Cependant, aucune étude ne réunit à la fois conditions strictement anaérobies et utilisation substrat complexe non prétraité en absence de toute autre source de carbone. De plus, alors que dans la nature les écosystèmes digestifs sont parmi les meilleurs pour dégrader la lignocellulose, ceux-ci n'ont jamais été testés pour leur capacité à conserver leur fonction dans des processus de fermentation industrielle.

Dans ce chapitre est décrite la stabilisation en fermenteur d'une communauté microbienne active sur lignocellulose, à partir de rumen bovin comme inoculum. La méthode utilisée est un enrichissement par cultures séquentielles en présence de paille de blé comme unique source de carbone. Une fois une communauté stable d'une culture à l'autre obtenue, ses capacités et sa cinétique de dégradation ont été caractérisées en suivant l'évolution de la dégradation et de la production d'acides gras volatils au cours d'un cycle de fermentation.

II. Efficient anaerobic transformation of raw wheat straw by a robust cow rumen-derived microbial consortium

Ce chapitre a été publié dans la revue « Bioresource Technology » sous le titre :

Efficient anaerobic transformation of raw wheat straw by a robust cow rumen-derived microbial consortium.

Lucas Auer[a,b,c,1], Adèle Lazuka[a,b,c,1], , Sophie Bozonnet[a,b,c], Diego P. Morgavi[d], Michael O'Donohue[a,b,c], Guillermina Hernandez-Raquet[a,b,c,*]

[1)] These authors contributed equally to this work
[a)] Université de Toulouse, INSA, UPS, LISBP, 135 Avenue de Rangueil, F-31077 Toulouse, France
[b)] INRA, UMR792, Ingénierie des Systèmes Biologiques et des Procédés, F-31400 Toulouse, France
[c)] CNRS, UMR5504, F-31400 Toulouse, France
[d)] INRA, UR1213 Herbivores, Centre de Theix, F-63122 St-Genès-Champanelle, France

II.1. Abstract

A rumen-derived microbial consortium was enriched on raw wheat straw as sole carbon source in a sequential batch-reactor (SBR) process under strict mesophilic anaerobic conditions. After five cycles of enrichment the procedure enabled to select a stable and efficient lignocellulolytic microbial consortium, mainly constituted by members of Firmicutes and Bacteroidetes phyla. The enriched community, designed rumen-wheat straw-derived consortium (RWS) efficiently hydrolyzed lignocellulosic biomass, degrading 55.5% w/w of raw wheat straw over 15 days at 35 _C and accumulating carboxylates as main products.

Cellulolytic and hemicellulolytic activities, mainly detected on the cell bound fraction, were produced in the earlier steps of degradation, their production being correlated with the maximal lignocellulose degradation rates. Overall, these results demonstrate the potential of RWS to convert unpretreated lignocellulosic substrates into useful chemicals.

II.2. Introduction

Lignocellulose (LC) is the most abundant terrestrial reservoir of renewable carbon on Earth and thus its use as feedstock for industrial processes is considered to be a viable alternative to fossil resources. In the field of bioconversion, the production of ethanol, methane and H2 from LC has been widely studied (Chang et al., 2010; Yan et al., 2012). Similarly, carboxylates (volatile fatty acids – VFAs) are also an interesting product of LC bioconversion, since these are intermediate building blocks for the production of polyhydroxyalkanoates, which are potentially useful as bio plastics (Kleerebezem and van Loosdrecht, 2007). Nevertheless, whatever the target molecule, much progress remains to be achieved to overcome the recalcitrant nature of LC.

Lignocellulosic biomass is a composite material, with cellulose (30–55%), hemicelluloses (18–37%) and (10–30%) lignin as the main constituents. These macromolecules are organized in a complex, heterogeneous matrix that confers to LC hydrophobicity, strength and resistance to biotic and abiotic attack.

In Nature, the recycling of LC is performed by LC-utilizing microbial consortia, present in soils and guts of animals, which deploy complex arsenals of enzymes including, cellulases, hemicellulases and lignin-degrading enzymes to overcome the complexity of LC. Accordingly, microbial consortia constitute an interesting paradigm for LC bioconversion. In particular, the use of microbial consortia in controlled bioreactors is a promising alternative to the deployment of single microorganisms. In contrast to the latter, microbial consortia do not require sterile operating conditions, display a more diversified enzymatic arsenal, show high metabolic diversity and versatility and generally remain stable and robust in a broad range of conditions (Kleerebezem and van Loosdrecht, 2007). Moreover, microbial consortia are more advantageous than pure strains artificially mixed in the laboratory because they rely on natural symbiotic interactions and coupled metabolisms that provide the basis of their robustness and can prevent undesirable phenomena in bioreactors, such as feedback inhibition. Overall, assuming that the advantages of microbial consortia can be fully harnessed in artificial environments, these constitute an interesting biocatalytic alternative for the industrial bioconversion of LC into useful products.

In previous studies, microbial consortia operating under both aerobic and anaerobic conditions have been successfully enriched on different types of biomass (Gao et al., 2014; Guo et al., 2010; Jiménez et al., 2014; Reddy et al., 2011; Trifonova et al., 2008). The selection of microbial consortia has been mainly directed towards the production of proteins

(Guo et al., 2010), or the biological pretreatment of LC and enhance methane production (Yan et al., 2012). Nevertheless, most of these studies used simplified cellulosic substrates (e.g. carboxymethylcellulose, Avicel, filter paper) or physico-chemically pretreated biomass (Eichorst et al., 2014; Yan et al., 2012), thus circumventing some of the difficulties linked to the complex structure of intact LC. Moreover, only a few studies have dealt with the enrichment of stable microbial consortia under strictly anaerobic conditions, thus the diversity and potential of microorganisms involved in LC decomposition in anaerobic environments remains poorly explored (Gao et al., 2014).

In this study, a natural bovine rumen consortium was enriched under strict anaerobic conditions, selecting for carboxylate production. Selection was performed by SBR cultivation, using raw wheat straw as the sole carbon source. Wheat straw was selected as substrate because it is considered one of the main LC feedstocks with about 34 Mt dry matter (DM) of extractible for being available in Europe for biorefinery purpose (Pajual and O'Donohue, 2013). Regarding the enrichment process, it is described herein in terms of biomass degradation, VFA production and also from the standpoint of the dynamic behavior of the microbial community. The major outcome of this study is the enrichment of a stable microbial community that displays good aptitude for the bioconversion of LC biomass into VFAs. Using the enriched consortium, the kinetic of LC deconstruction has been deeply studied, monitoring the physico-chemical composition of the substrate and characterizing the products. A particular effort was made to correlate enzyme deployment in both the liquid and solid fractions with biomass degradation to identify the rate limiting steps LC bioconversion process under strict anaerobic conditions.

II.3. Material and methods

II.3.1 Lignocellulosic substrate and inoculum

20 kg of straw from the winter wheat variety Koreli was harvested at an experimental farm (INRA, Boissy-le-Repos, France) in August 2011. After harvest, the straw was milled to 2 mm and stored at room temperature (20–25°C). The non-producing Holstein dairy cows used in this study were reared according to the national standards fixed by the legislation on animal care (Certificate of Authorization to Experiment on Living Animals, No. 004495, Ministry of Agricultures, France). In particular, the cows were fed a standard dairy cow ration composed of corn silage (64% DM), hay (6% DM) and concentrate (30% DM). They were fed ad libitum once a day in the morning and sampling was done just before feeding. For sampling, the whole rumen content was collected just before feeding from two individuals fitted with rumen cannula. Rumen contents were taken from various parts of the rumen and manually homogenized. Subsamples (30 g) were immediately frozen in liquid nitrogen before storage at -80°C.

II.3.2 Anaerobic reactors

Bioconversion was carried out in anaerobic SBR (2L BIOSTAT A+, Sartorius, Germany) using a mineral medium containing wheat straw as the sole carbon source (20 $g.L^{-1}$). The mineral medium (MM) contained, per liter of distilled water: KH_2PO_4, 0.45 g, K_2HPO_4, 0.45 g; NH_4Cl, 0.4 g; NaCl, 0.9 g; $MgCl_2$-$6H_2O$, 0.15 g; $CaCl_2$-$2H_2O$, 0.09 g. MM was supplemented with 250 µL of V7 vitamin solution (Pfennig and Trüper, 1992), and 1 mL trace elements solution, containing per liter of distilled water: H_3BO_3, 300 mg; $FeSO_4$-$7H_2O$, 1.1 g; $CoCl_2$-$6H_2O$, 190 mg; $MnCl_2$-$4H_2O$, 50 mg; $ZnCl_2$, 42 mg; $NiCl_2$-$6H_2O$, 24 mg; $NaMoO_4$-$2H_2O$, 18 mg; $CuCl_2$-$2H_2O$, 2 mg; sterilized by filtration (0.2 µm).

Reactors were operated under mesophilic temperature (35°C) and pH was regulated at 6.15 by the addition of 1 M NaOH. Anaerobic conditions were ensured by nitrogen flush at the beginning of the experiment. To avoid methanogenic conditions bromoethansulfonate (BES), a methanogenesis inhibitor, was added at the beginning of experiment (1 mM). During the experiment, if methane was detected in the biogas, BES was spiked until a maximum concentration of 10 mM.

To progressively enrich a LC-degrading microbial consortium, a first batch reactor was initiated by inoculating (2% w/v) fresh medium with raw cow rumen (pooled from the two cows, thawed overnight at 4°C). This reactor was operated for 7 days before removing 10% v/v of its content, which was used as inoculum for the subsequent batch reactor. In this manner, SBR cultures were performed until VFA reached a stable concentration. Thereafter, 200 mL samples of the enriched consortium RWS were frozen under liquid nitrogen and stored at -80°C. After thawing at 4°C overnight, these frozen samples of the RWS consortium were used as inocula for further experiments. Deeply characterization of LC degradation by the enriched RWS consortium was performed in three biological replicate reactors incubated for fifteen days in the conditions described above. Reactors were carried out at three months intervals; they were inoculated with the same source of inoculum stored at -80°C.

II.3.3 Chemical analyses

Total solids (TS) were measured using 10 mL samples that had first been centrifuged (7197xg, 10 min), rinsed twice with distilled water and dried 24 h at 105°C. The mineral fraction (MF) was estimated by mineralization of the samples at 500°C for 2 h, and volatile solids (VS) were estimated from the difference between TS and MF. VS degradation was expressed as weight/weight percentages. Wheat straw composition was determined using the sulfuricacid hydrolysis method described by de Souza et al. (2013) on 40 mg samples. Quantitative detection of monomeric sugars was achieved using high performance liquid chromatography (HPLC) on a Ultimate 3000 Dionex separation system equipped with a BioRad Aminex HPX 87H affinity column and a refractive index detector (Thermo Scientific) as described by Monlau et al. (2012).

Fourier transform infrared spectroscopy (FT-IR) analysis was performed on lyophilized samples using an attenuated total reflection (ATR) Nicolet 6700 FT-IR spectrometer (Thermo Fisher) equipped with a deuterated-triglycine sulfate (DTGS) detector. Spectra were recorded in the range 400-4000 cm^{-1} with a 4 cm^{-1} resolution. Each sample was recorded 5 times and analyses were performed using average spectra resulting from 32 individual scans. For FT-IR spectral analysis, the peak ratio 1512:1375 cm^{-1} was considered to be a lignin:holocellulose ratio (Monlau et al., 2012).The peaks 1430 cm^{-1} and 898 cm^{-1} were attributed to the amounts of crystalline and amorphous cellulose respectively (Monlau et al., 2012) and the ratio of these peaks was considered to be the lateral order index (LOI),

which is the ratio of crystalline: amorphous cellulose. VFA production was determined by gas chromatography (GC), using a Varian 3900 chromatograph as described by Cavaillé et al. (2013). The total organic carbon (TOC) content of the liquid fraction was measured using a TOC analyzer (TOC-VCSN, Shimadzu Co., Japan). Gas composition was analyzed using a chromatograph HP 5890 equipped with a conductivity detector and a HAYSEP D column. All the macro-kinetic parameters are expressed as average values obtained in triplicate biological reactors.

II.3.4 Enzyme activity assays

For enzyme activity measurement, triplicate reactor samples (5 mL) were removed at regular intervals. Samples were centrifuged at 7197xg for 10 min at 4°C thus yielding a supernatant and a solid pellet. It was assumed that the supernatant contains extracellular enzymes, while the pellet is representative of cell-bound enzymes. The latter was suspended in 6 mL of acetate buffer solution (50 mM, pH 6), placed on ice and sonicated four times for 20 s each using a sonication device (Bandelin Sonoplus HD 2070, MS73 probe) operating at 60 W.

For each reactor and each sampling time, end point enzymatic activities were measured in technical duplicates in both the extracellular and cell-bound fractions. Enzymatic activities were expressed as average values obtained on triplicate reactors.

Xylanase and endoglucanase (CMCase) activities were measured using 1% w/v xylan beechwood (Sigma) and 1% w/v carboxymethyl cellulose (CMC) (Sigma) respectively dissolved in buffer solution. Activities were estimated by measuring the release of reducing sugar equivalents after incubation at 35°C for 1 h (xylanase) or 4 h (CMCase). Reducing sugars were determined using the dinitrosalicylic acid (DNS) method, measuring absorbance of the sample at 570 nm using a UV/VIS spectrophotometer (Multiskan Ascent, Thermo Scientific). One unit of CMCase or xylanase activity (UA, unit of activity) was defined as the amount of enzyme that produces 1 µmol of reducing sugars per minute. Exoglucanase (cellobiohydrolase) and β-glucosidase activities were assessed using fluorescent methylumbelliferyl (MUF) substrates (MUF cellobioside and MUF glucopyranoside, Sigma) and methylumbellipherone as standard. Incubation was carried out for 2 min in the same conditions as above. Measurements were made with a Cary Eclipse Fluorescence Spectrophotometer (Agilent Technologies, Santa Clara, CA, USA). One unit of exoglucanase

or β-glucosidase activity (UA) was defined as the amount of enzyme that produces 1 µmol of methylumbellipherone per minute.

II.3.5 Analysis of microbial diversity

For bacterial diversity analysis, samples (1.5 mL) were removed and immediately centrifuged (13,000xg, 5 min, 4°C). Subsequently, the supernatant was removed and the solid pellet was frozen in liquid nitrogen and stored at -80°C. Simultaneous DNA/RNA extraction was performed using the PowerMicrobiome RNA isolation kit (MoBio Laboratories, Carlsbad, CA, USA) following the manufacturer's instructions. Cell lysis was carried out using a FastPrep Instrument (MP Biomedicals, Santa Ana, CA, USA; 2x30 s at 4 ms^{-1}). Subsequently, DNA purification was carried out using the AllPrep DNA/RNA mini kit (Qiagen), following the manufacturer's instructions. Microbial diversity was analyzed using MiSeq Illumina sequencing performed by the GenoToul Genomics and Transcriptomics facility (Auzeville, France). A 460 bp fragment corresponding to the highly variable V3–V4 region of 16S rRNA gene was amplified from genomic DNA samples using bacterial universal primers F343 and R784 (Table 1). Primer sequences and PCR conditions are detailed in Table 1. The resulting PCR products were purified and loaded onto the Illumina MiSeq cartridge according to the manufacturer's instructions. Joined-pair reads (56,000 reads per sample, minimum 30,224) were pre-processed and filtered for chimeras using a Mothur pipeline (Schloss et al., 2009).

Operational taxonomic units (OTU; 97% similarity) were defined using a random subsample of 15,000 sequences per sample using Mothur default parameters, and OTUs that were less than 10 sequences across all samples were filtered. OTUs were taxonomically classified using Mothur and LTPs115 SSU database. Comparative analysis of samples was performed on a random subsample of 10,000 high quality sequences per sample. These produced saturated rarefaction curves (Supplementary data S. 1) and were sufficient to assess relationships among samples (Caporaso et al., 2011).

Table 1
PCR primers and conditions.

PCR1	F343	5′-CTTTCCCTACACGACGCTCTTCCGATCTTACGGRAGGCAGCAG-3′
	R784	5′-GGAGTTCAGACGTGTGCTCTTCCGATCTTACCAGGGTATCTAATCCT-3′
	Cycles	94 °C/60 s – 30 × (94 °C/60 s – 65 °C/60 s – 72 °C/60 s) – 72 °C/10 min
PCR2	FP2	5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC-3′
	RP2	5′-CAAGCAGAAGACGGCATACGAGAT-**index**-GTGACTGGAGTTCAGACGTGT-3′
	Cycles	94 °C/60 s – 12 × (94 °C/60 s – 65 °C/60 s – 72 °C/60 s) – 72 °C/10 min

Community diversity was estimated using the Simpson and Shannon diversity index. The richness, at 97% similarity, was estimated using Chao and the abundance coverage estimator (ACE). The distance between communities present in different samples was estimated using the Bray–Curtis dissimilarity index and weighted Unifrac distances, calculated using the Phyloseq R package v1.6.1 (McMurdie and Holmes, 2013). These distances were used to ordinate samples with Hclust R function. The sequencing data generated in this study were deposited in the NCBI Sequence Read Archive and are available under the project number PRJNA284586.

II.4. Results and discussion

II.4.1 Enrichment: macro-kinetic and diversity analysis

In this study, a sequential culture enrichment strategy was deployed to obtain a stable rumen-derived microbial community displaying good LC biomass hydrolysis capability. Likewise, the aim was to reduce the complexity of the initial microbial inoculum and to study the resultant LC-degrading, VFA-producing microbial community.

Gratifyingly, after only 10 SBR subculture cycles a stable consortium (RWS) was obtained that is characterized by its ability to release up to 35% VS, accumulating up to 1.94 ± 0.04 gC VFA L^{-1} over a 7-day period when operating at 35°C (Fig. 1a). Amplicon sequencing of the 16S rDNA gene revealed that important changes in the microbial community had accompanied the acquisition of higher VFA production and LC biomass degradation capability (Fig. 1b). The majority of these changes occurred in the initial enrichment phase, specifically by the end of the second SBR cycle. These radical changes can almost certainly be attributed to the transition from the natural environment of the rumen to the artificial conditions of the bioreactor and thus mainly correspond to community

acclimation. Compared to the rumen inoculum, the community richness, estimated by the ACE and Chao methods, decreased strongly from 282.9 and 287.4, respectively, to 241.5 and 255.1 after the first SBR cycle and to 203.7 and 205.6 at the second SBR cycle. Similarly, determination of the Simpson diversity and Shannon index revealed a loss in diversity, which was particularly marked during the two firsts SBR cycles and then stable along the subsequent enrichment cycles (Table 2). Comparison of the Simpson diversity across the whole experiment, from the rumen-derived inoculum to the final SBR cycle (C10), revealed that it had dropped from 36.8 to 11.7. Similarly, the Shannon index decreased from 4.45 to 3.38. Hierarchical Unifrac clustering of the initial community presented in the inoculum and those generated at the end of each SBR cycle revealed that two community groups cluster together (Fig. 1c). The first cluster encompasses the communities present during the first four SBR cycles, whereas the second one contains the communities from SBR cycles 5 to 10.

Correlation of these results with the higher degradation of LC biomass and VFA production (described below) suggests that the microbial community evolved towards a structure that better expresses the wheat straw degradation and VFA production functions within the conditions defined by the experiment. Significantly, the communities that grouped in the second cluster are rather similar, reflecting the relative stability of the community's structure after five cycles of SBR enrichment. Similar rapid stabilization of enriched microbial communities has already been reported by Jiménez et al. (2014) and Trifonova et al. (2008), with stable microbial communities being obtained after six transfers on wheat straw and five transfers on grass fiber respectively. Overall, this rapid stabilization underlines the stringency of the enrichment conditions and suggests that the use of LC biomass as a sole carbon source is a good way to select a LC-degrading microbial community.

Table 2
Variation of richness (ACE and Chao) and diversity index (Simpson and Shannon) through the enrichment process.

SBR cycle	Ti	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Shannon index	4.45	3.71	3.25	3.51	3.47	3.04	3.11	3.16	3.17	3.08	3.38
Simpson index	36.8	21.1	12.6	16.1	14.9	7.7	7.8	8.9	9.8	8.2	11.7
ACE richness	282.9	241.5	203.7	219.7	195.5	191.4	199.0	192.4	195.5	191.7	186.5
Chao richness	287.4	255	205.6	224.7	194.2	191.3	201.3	193	217.8	185.5	187.3

Ti: initial time. C*n*: cycle number.

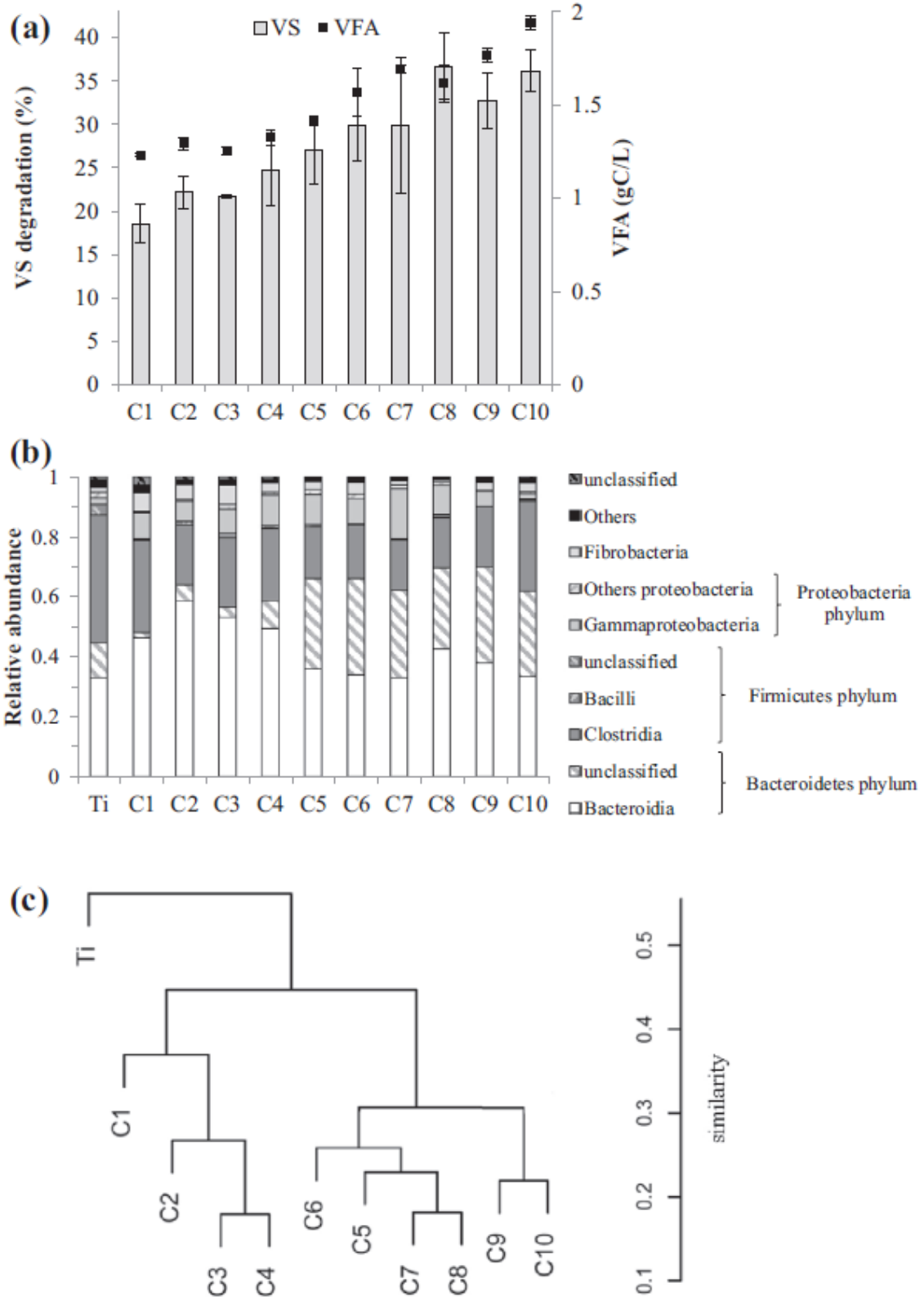


Figure 1: VS degradation and VFA production (a), relative community composition (b) and community clustering (c) through the enrichment procedure. Error bars in VS degradation are standard deviations of two technical replicates.

The analysis of the community's composition during the successive SBR cycles showed a dominance of OTUs related to the phylum *Bacteroidetes* (44-70%), followed by *Firmicutes* (17-46%) and *Proteobacteria* (2-18%), which constituted more than 85% of the total community (Fig. 1b). The relative abundance of the *Bacteroidetes* phylum increased during the first five SBR subcultures (from 44% to 66%); and stabilized between the fifth and the last cycle of enrichment (Fig. 1b). Within the *Bacteroidetes* phylum, it is noteworthy that the *Bacteroidia* class diminished during the enrichment process to the benefit of an unclassified class, composed of only one OTU. This latter was related to the unclassified *Bacteroidetes* clone P5_J07 (EU381760, 100% identity) that has already been identified in the ruminal microflora (Kong et al., 2010). This OTU increased from a relative abundance of 3% in the inoculum to 28% in the SBR cycles 5-10. A second dominant OTU that was enriched during the SBR subculture process (from 1% to 14%) was closely related to *Bacteroides graminisolvens* strain JCM 15093 (*Bacteroidetes* phylum, 100% identity), a known xylanolytic anaerobic bacterium able to produce acetate, propionate and succinate (Nishiyama et al., 2009). *Prevotella* genus was well represented with several OTUs including *Prevotella ruminicola* and *Prevotella brevis*, known ruminal bacteria able to produce succinate that can be further decarboxylated into propionate (Ramsak et al., 2000); OTUs related to unclassified Prevotella species were also present. The *Clostridia* class (*Firmicutes* phylum), highly abundant in the rumen inoculum (43%), decreased after the first cycle of SBR enrichment, but then stabilized at a relative abundance of 20% in subsequent SBR cycles. This class included *Butyrivibrio* species, able to transform xylan and produce butyrate (Krause et al., 2003); *Ruminococcus albus*, a ruminal cellulose degrader and acetate producer (Christopherson et al., 2014); various *Clostridium* species, reported as displaying cellulase and/or hemicellulose activities; *Eubacterium cellulosolvens*, a cellulose degrader known to produce mainly butyrate, succinate and valerate (Prins et al., 1972), as well as numerous unclassified *Clostridiales*. Therefore, it appears that members of RWS consortium are particularly well-equipped to degrade LC biomass and produce VFA. These observations are consistent with those of previous studies that have identified *Firmicutes* and *Bacteroidetes* as dominant phyla in LC-degrading ecosystems and are known to participate in the hydrolytic and acidogenic steps in carboxylate production systems (Hollister et al., 2011). Moreover, the enrichment of members of *Bacteroidetes* has been previously observed in experiments where strict anaerobic conditions have been applied (Gao et al., 2014). During the enrichment process the hydrolytic capability of the microbial community was estimated by measuring the residual VS content

and metabolite production. This revealed that biomass degradation almost doubled between the first SBR cycle and the final enrichment step, increasing from 20% to 35% VS (Fig. 1a). This increase in biomass degradation was particularly marked during the first four SBR cycles (from 18.5 ± 2.2% VS to 24.7 ± 4.1% VS), after which progress was steadier (between 24.7 ± 4.1% VS and 36.1 ± 2.4% VS). Similar behavior was observed on VFA accumulation, increasing from 1.23±0.01 to 1.56±0.12 $gC.L^{-1}$ between the first and sixth cycles (Fig. 1a). Significantly, stabilization of the community's LC biomass-degrading function correlates quite well with the evolution of the community's composition, since from the fifth SBR cycle the community grouped together with the final one (Fig. 1c). This result clearly underlines the benefit of the enrichment procedure and the fact that only ten successive cultures were necessary to procure a functionally- and microbiologically-stable consortium. Moreover, from the fifth to the tenth cycle of enrichment, the LC degradation ability and community composition were stable and reproducible after each cycle of culture, showing the robustness and reproducibility of the selected microbial community. These observations strongly suggest that the LC biomass degradation function, concomitant to VFA production, is linked to the presence of specific species. In this respect it is useful to recall that during the enrichment process the microbial community became progressively enriched in OTUs related to the *Bacteroidetes* phylum and the *Clostridia* class; some of them known for their capacity to produce VFA (see Supplementary data S.2).

Previous studies focused on the production of stable consortia presenting high LC biomass degradation capability, already demonstrated how progressive enrichment of inocula (e.g. compost, soil, and sediment) can be used to procure microbiologically-stable LC biomass-degrading communities, but did not always evidence increases in the potency of the LC biomass-degrading function (Gao et al., 2014; Guo et al., 2010). However, working with a bovine rumen inoculum (1% v/v), Chang et al. (2010) did obtain results that are consistent with those described here. However, in this latter work a much longer enrichment process (18 cycles) was necessary to obtain a stable community, composed mainly of Clostridia. Such community was able to release 27% w/w of hemicellulose and 2% w/w of cellulose after 8 days of cultivation on milled (1 mm) Napiergrass (1.5% w/v); whereas in the present study degradation was much higher (35.5% w/w in 7 days) after only 10 cycles of enrichment. Similarly, the aerobic enrichment of a compost-derived community on switchgrass and corn stover procured very fast increases in LC biomass-degrading function, with degradation

increasing from 8% to 34% (switchgrass) and from 19% to 23% (corn stover) in only 4 cycles (Reddy et al., 2011). Remarkably, in this study it was shown that the main changes occurred between the first and second enrichment cycles, since thereafter both the function and the community structure remained quite stable. Taken together with results obtained in the present study, it is clear that enrichment can be an efficient method to select for LC biomass-degrading function with strong selection occurring after only a relatively small number of subculture steps. However, the efficiency of the process is almost certainly determined by the biological potential of the initial inoculum, the culture conditions and the specific composition of the biomass. To the authors' knowledge, this study represents the first work demonstrating anaerobic enrichment of rumen-derived inoculum on raw wheat straw as an efficient tool to increase substrate degradation capability and generate a stable and robust consortium.

II.4.2 Characterization of wheat straw degradation by consortium RWS

.II.4.2.1 Wheat straw degradadation and VFA production by RWS

The LC degradation kinetics of the enriched microbial consortium RWS, obtained after ten SBR subcultures, was studied in detail using the same culture conditions as those used for enrichment. The inoculation rate was 10% v/v and incubation was pursued over a 15-day period. After 7 days, % VS degradation was 36.8±3.6 (Fig. 2a), reaching 55.5±2.8 after 15 days. Simultaneously, VFA production reached 2.92±0.45 $gC.L^{-1}$. It is important to note that, although the reactors were not performed simultaneously and that the same source of inoculum was stored for different periods (as detailed in Methods), similar LC degradation and VFA production rates were observed in all replicate reactors. Taken together with the remarkable stability of the consortium observed over SBR cycles 5 –10, it is clear that RWS is highly robust. The study of the kinetics of VS release and VFA production revealed a 3-phase dynamic behavior. VS degradation was rather low at the beginning of the experiment and then strongly increased from days 2 to 10. Consistently, before day 2, average VFA productivity was 0.155±0.027 $gC_{VFA}.L^{-1}.day^{-1}$, increasing thereafter to 0.224±0.008 $gC_{VFA}.L^{-1}.day^{-1}$. However, from day 10 onwards, both VS degradation and VFA productivity declined (0.110±0.100 $gC_{VFA}.L^{-1}.day^{-1}$). TOC measurements (data not shown) performed on the supernatants showed that no other metabolites (other than VFA) were present. The VFA molar ratio (expressed as %) for acetic (AA), propionic (PA), butyric (BA) and valeric (VA) acid was 65.6±3.9%, 20.5±3.4%, 10.0±2.9% and 3.8±3.1% respectively, with this composition remaining rather constant during the 15-day period (Fig. 2b). This VFA profile is

quite typical for acidogenic systems at neutral and slightly acidic pH (Hu and Yu, 2005). Similarly, the CO2:VFA ratio also remained constant (0.33±0.04 gC_{CO2} / gC_{VFA}), with CO2 production representing approximately 25% (gC/gC) of all products. Using the theoretical stoichiometry of the bioconversion of glucose to C2-C4 VFA (AA, PA, BA) in ruminant digestive systems (Eqs. (1–3)), and assuming that valeric acid production results from the condensation of acetyl-CoA and propionyl-CoA with oxidation of reduced cofactors (Eq. (4)), the yields of AA, PA, BA and VA can be calculated as 0.66, 0.82, 0.48 and 0.56 $g.g^{-1}$ respectively (Hu and Yu, 2005; Nozière et al., 2010). Moreover, calculation of the theoretical stoichiometry of VFA production from C6 degraded sugars based on the wheat straw composition determined in this study reveals a yield of 0.65 $g.g^{-1}$ (or 0.74 g eq AA g^{-1}). However, the experimentally-measured VFA yield was 0.49 g g_1 (or 0.61 g eq AA g^{-1}), which represents 75% of the theoretical yield. This high VFA yield was achievable because of anaerobic conditions, since unlike aerobic conditions these limit carbon loss in the form of CO2 and microbial biomass (Feng et al., 2011). It is noteworthy that the yield of VFA observed in this experiment is quite consistent with that observed in previous studies (0.6-0.7 g eq AA g_1 VS) in which rumen inocula were used to treat corn stover (Datta, 1981; Hu and Yu, 2005).

$$C_6H_{12}O_6 + 2H_2O \rightarrow 2CH_3COOH + 2CO_2 + 4H_2 \quad (1)$$

$$C_6H_{12}O_6 + 2H_2 \rightarrow 2CH_3CH_2COO_H + 2H_2O \quad (2)$$

$$C_6H_{12}O_6 \rightarrow 2CH_3CH_2CH_2COOH + 2CO_2 + 2H_2 \quad (3)$$

$$C_6H_{12}O_6 + 2H_2 \rightarrow CH_3CH_2CH_2CH_2COOH + CO_2 + 2H_2O \quad (4)$$

Using the same equations, the theoretical CO2 yield is 0.35 $g.g^{-1}$, while that measured in the present experiments was 0.26±0.03 $g.g^{-1}$ biomass-derived sugars. The discrepancy between the theoretical and experimentally-measured yields is probably the result of the utilization of a part of the carbon for microbial growth and maintenance, but could also result from CO2 consumption via homoacetogenic reactions. However, although increased quantity of DNA clearly indicated that microbial growth did occur during the experiment (data not shown), it was not possible to estimate quantitative microbial growth using standard methods, because of interference from the solid wheat straw fraction.

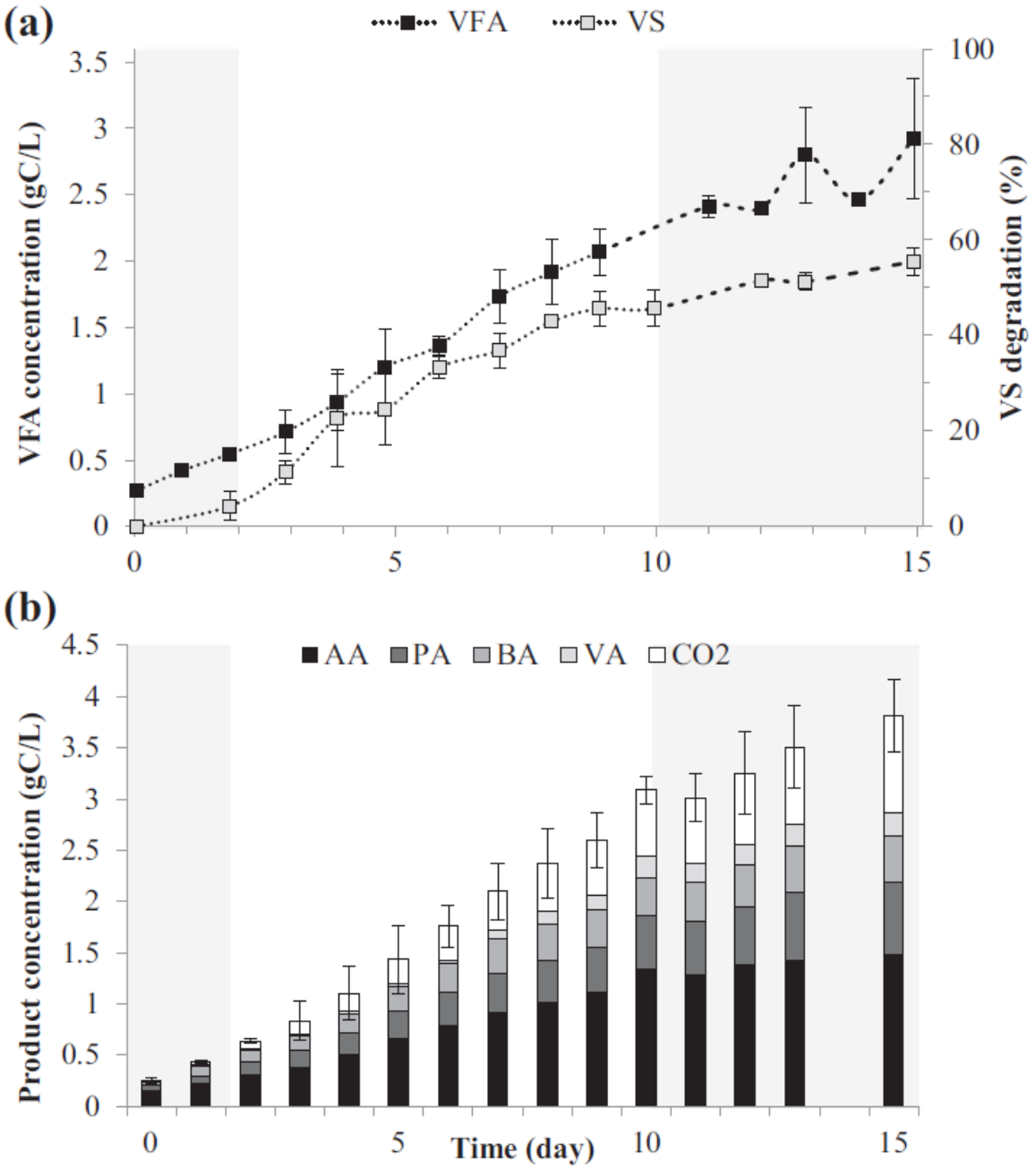


Figure 2: Profile of VS degradation and VFA production (a) and products composition of RWS cultivated on wheat straw (b). Error bars are standard deviations of three biological replicates.

.II.4.2.2 Effects of RWS activity on the physicochemical composition of wheat straw

The residual LC biomass was characterized using biochemical (hydrolysis and HPLC) and physical (FT-IR) methods. The monitoring of cellulose and hemicellulose content, performed at several points during the experiment, revealed that these plant polysaccharides were simultaneously degraded, with their hydrolysis following the same kinetic profile and displaying similar residual percentages (Fig. 3a). This behavior was also similar to the VS

degradation kinetics, displaying a high degradation between days 2 and 10, and a relative stability beyond day 10. Thus, the RWS consortium did not preferentially degrade one or the other of these fractions. At the end of the 15-day incubation, more than 60% w/w of holocellulose was degraded. In contrast, the determination of lignin content revealed that this macromolecule was not degraded (data not shown).

Monitoring wheat straw degradation by RWS using FT-IR analysis revealed an increase in both the ratio of the 1512:1375 cm^{-1} peaks (representing the lignin:holocellulose ratio) and the LOI, representing the ratio of crystalline:amorphous cellulose (Fig. 3b). Although these measurements are associated with quite large standard deviations, the time-dependent profile nevertheless correlates well with the 3-phase dynamic behavior of VS loss

(Fig. 2a) and the evolution of the two ratios is opposite to that of the residual holocellulose measurements (Fig. 3a). Therefore, accounting for the fact that lignin was not degraded, the increase of 1512:1375 cm^{-1} ratio is consistent with a decrease of the holocellulose. Moreover, the increase of the LOI ratio indicates that the RWS consortium preferentially degraded the amorphous cellulose fraction, leaving a residual cellulose fraction enriched in crystalline cellulose.

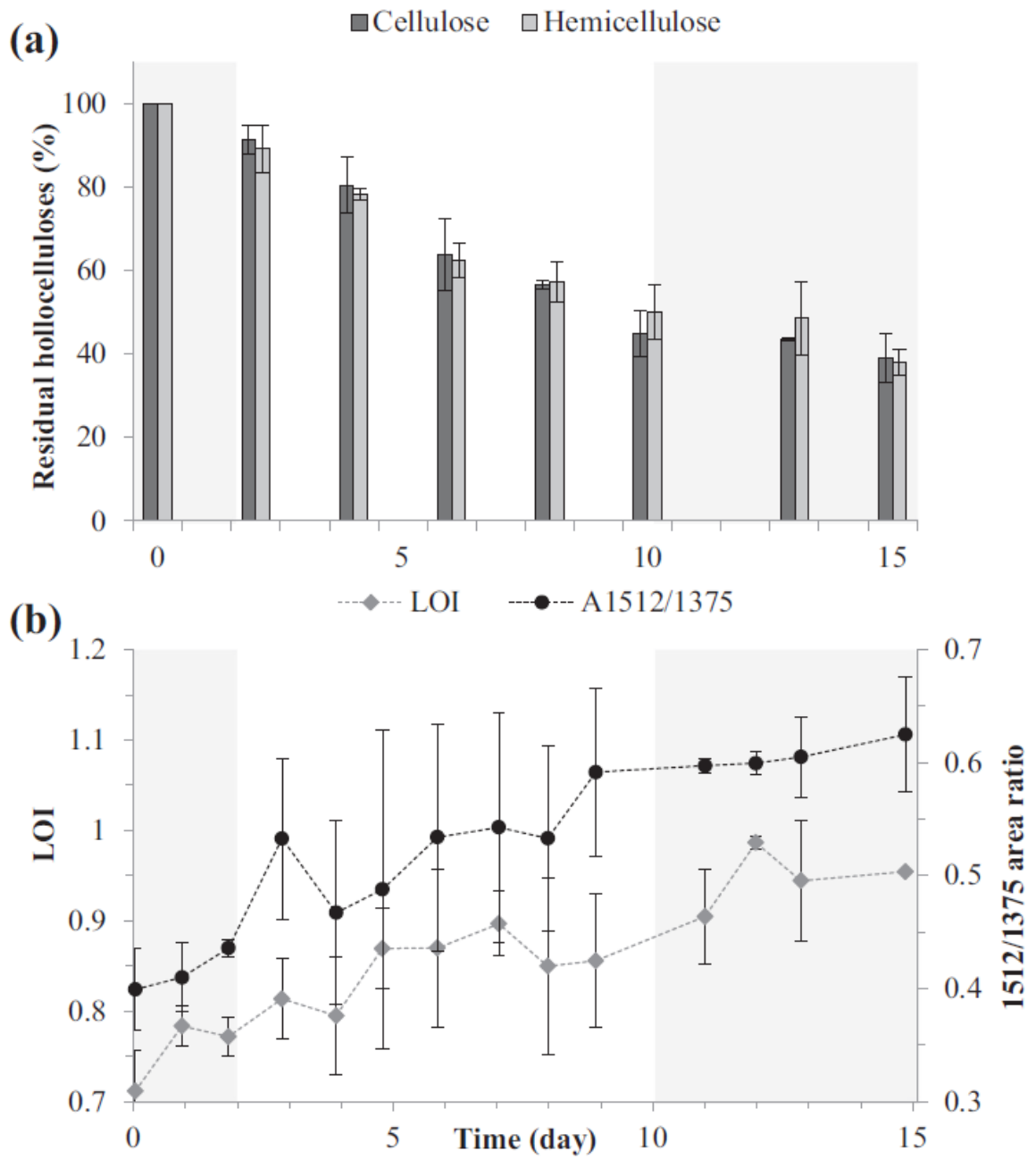


Figure 3: Substrate characterization: residual percentages of cellulose and hemicellulose calculated as percentage from their respective initial concentrations (a) and lateral order index (LOI) and 1512:1375 cm^{-1} area ratio (b). Error bars are standard deviations of three biological replicates.

So far, only a few studies involving the use of enriched microbial consortia have provided detailed insight into how the LC biomass is degraded during the process. Using a compost-derived consortium (XCD-2) enriched on alkali-pretreated lignocellulosic substrates, Guo et al. (2010) showed that LC biomass degradation mainly involved the destruction of the

hemicellulosic fraction (up to 89% w/w), with cellulose degradation being only about 12% w/w. This is in sharp contrast to the present study, but the fact that a rumen-derived consortium and untreated raw LC biomass were used may well have orientated the microbial enrichment process, leading to a more efficient functional community. Indeed, alkaline treatment leads to saponification of ester bonds that link hemicellulose to lignin and decorating side-chain groups (e.g. acetyl and feruloyl moieties) to the main xylan chain. The overall effect is to increase the solubility and accessibility of hemicellulose and thus provide a prime target for microbial attack (Hendriks and Zeeman, 2009).

The finding that RWS preferentially degrades the amorphous fraction of cellulose is logical, since it is known that the crystalline fraction is quite resistant to the action of microbial enzymes (Lynd et al., 2002). The absence of lignin degradation is consistent with the fact that anaerobic bacteria are not widely associated with lignin degradation, this function being much more frequently associated with aerobic fungi, which produce oxidative enzymes such as peroxidases and laccases (Ten Have and Teunissen, 2001). Moreover, studies on bacterial consortia incubated on unpretreated LC substrate (switchgrass) under aerobic conditions, where lignin modification by peroxidase and lacasse oxidation-reactions likely occurs, showed that the physicochemical composition of lignin remain unchanged in the residual biomass after two weeks of incubation (Eichorst et al., 2014). It is clear that native lignin in raw biomass is resistant to bacterial attack. Nevertheless, despite the fact that lignin was not degraded by RWS, the consortium was able to efficiently degrade holocellulose, demonstrating that lignin modification is not a prerequisite feature of biomass deconstruction by microbial consortia under the applied experimental conditions. The ability of RWS to concomitantly degrade hemicellulose and cellulose fractions of raw LC substrate to produce VFA is a clear advantage that may have led to the higher LC degradation observed in the present study.

.II.4.2.3 Enzymatic activity profiles along wheat straw degradation

The enzyme activity (xylanase and CMCase) profiles expressed by RWS during the 15-day experiment displayed an initial lag phase that covered the first 2 days (Fig. 4a). Afterwards, from day 2 to 7 xylanase and CMCase activities increased 4- and 2.5-fold respectively, reaching maximum levels of 0.81±0.11 UA mL^{-1} and 0.04±0.01 UA mL^{-1}. Towards the latter part of the experiment CMCase and xylanase activities stabilized. In sharp

contrast, exoglucanase and b-glucosidase activities decreased throughout the 15-day period, even if there was a slight increase in exoglucanase activity between days 5 and 6 (Fig. 4b). Taken together, these results imply that an endocellulase-like activity (i.e. CMCase) played a more important role in LC biomass degradation by RWS than an exocellulase activity. This is consistent with the fact that amorphous cellulose was the main target for RWS, since it is known that endocellulases mainly act on amorphous regions, whereas exocellulases preferentially act on crystalline cellulose (Lynd et al., 2002).

The localization of the different enzyme activities that were produced by RWS during the experiment revealed that the different enzyme activities were mainly localized in the insoluble pellet fraction (Fig. 4c). This might imply that different members of the RWS community produced cell-bound enzymes that are perhaps part of cellulosome-like complexes described by Lamed et al. (1983) and observed by Gao et al., (2014). Indeed, several *Clostridia* members known for having a cellulosome complex, such as R. albus, were detected in RWS (Christopherson et al., 2014). Nevertheless, the fact that both microbial biomass and wheat straw were present in the pellet fraction, it is impossible to completely affirm that all of the enzymes were actually cell-bound. This is because some free, extracellular enzymes could have remained associated with the solid biomass perhaps forming enzyme-substrate complexes involving high affinity carbohydrate binding domains (Himmel et al., 2010).

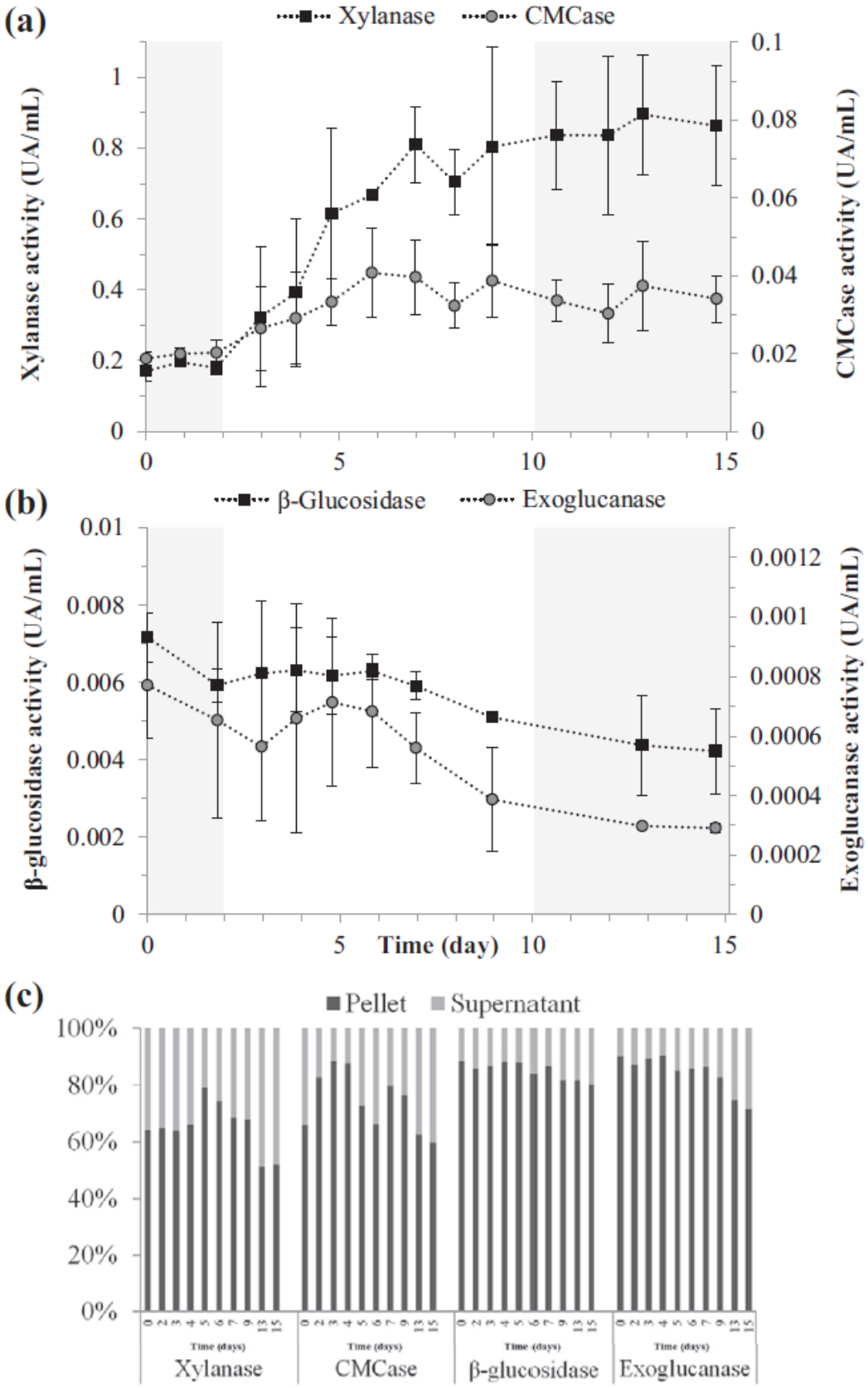


Figure 4: enzymatic activity profiles of RWS through wheat straw transformation: xylanase and CMCase activities (a); β-glucosidase and cellobiohydrolase activities (b) and pellet; supernatant repartition of measured enzymatic activities (c). Error bars are standard deviations of three replicates.

The integration of the different parameters (VS release, substrate characterization, VFA production, enzyme activities) that describe the degradation of wheat straw by RWS provides an overview of the whole experiment. An initial phase that covers the first two days was characterized by stable, low-level enzyme activity and VFA production and very little biomass degradation. During this phase VFA production was no doubt sustained by the degradation of readily-accessible biomass that could be easily dealt by only small amounts of enzyme. A second phase characterized by high enzyme activities, VFA production and biomass degradation was the result of the use of amorphous cellulose and hemicellulose, with this phase coming to end when the ability of the enzymes to furnish metabolizable sugars from the recalcitrant and more crystalline structure of the remaining substrate became overwhelmingly rate-limiting. Thereafter, despite maintenance of enzyme levels, substrate hydrolysis is low, a fact reflected by lowered VS degradation and VFA production. Overall, the present results clearly highlight the fact that enzyme hydrolysis and the increase of wheat straw recalcitrance throughout the experiment were the rate-limiting steps regulating of the wheat straw to VFA bioconversion process.

Among the studies of microbial consortia acting on LC biomass, only a few have so far simultaneously studied the dynamics of enzyme production and substrate modification, and even when enzyme activities have been studied, the cell-bound fraction has often been neglected (Hui et al., 2013). In the case of the xylanolytic consortium XDC-2, preferential degradation of the hemicellulose fraction was correlated with strong extracellular xylanase activity, with only a minor fraction of enzyme activity being cell-bound (Guo et al., 2010). However, it should be recalled that this experiment was subject to micro-aerobic conditions, unlike the conditions that prevailed in the present study. In a previous study performed under strict anaerobic conditions, cellulosome-like systems were detected and associated with single Clostridium-related isolates arising from the consortium SQD-1.1 (Gao et al., 2014). In this respect it is noteworthy that the consortium RWS was also characterized by numerous OTUs belonging to Clostridia, a class that is well-known for its cellulosomes. In view of the performance of RWS, it is obvious that further investigation of the enzyme array produced by RWS could be highly profitable route of investigation. Also, studies of RWS at metatrascriptomic and metaproteomic level are required to ultimately confirm the microorganisms and enzymes acting along wheat straw degradation.

II.5. Conclusion

This work reveals original insight into the dynamics and function of a microbial community acting on raw LC biomass and operating in non-sterile, strictly anaerobic conditions. The RWS enriched consortium transformed 55.5% w/w of the lignocellulose into VFA as main products. RWS is mainly constituted by members of *Bacteroidetes* and *Firmicutes* phyla and produced the key cellulolytic and hemicellulolytic enzymes, mainly of cell-bound type, to simultaneously degrade cellulose and hemicellulose fractions of raw biomass. The holistic analysis of substrate structure and composition, VFA and enzymes production underlines the potency of non-sterile microbial enrichment for the development of high-performance biomass-degrading consortia.

II.6. Acknowledgements

This research was supported by the French National Agency for Energy and Environment (ADEME), the Carnot Institute 3BCAR and the INRA Metascreen project. The authors thank the ICEO facility dedicated to enzyme screening and discovery, and part of the Integrated Screening Platform of Toulouse (PICT, IBiSA) for providing access to the Cary Eclipse Fluorescence spectrophotometer equipment. The authors also thank the Genomics and Transcriptomics (GET) INRA platform for their help with sequencing. M. Abadie, M. Bounouba and E. Mangelle are acknowledged for their assistance with experiments and technical support.

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.biortech.2015..07.084.

II.7. References

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., Fierer, N., Knight, R., 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc. Natl. Acad. Sci. U.S.A. 108, 4516–4522.

Cavaillé, L., Grousseau, E., Pocquet, M., Lepeuple, A.S., Uribelarrea, J.L., Hernandez Raquet, G., Paul, E., 2013. Polyhydroxybutyrate production by direct use of waste activated sludge in phosphorus-limited fed-batch culture. Bioresour. Technol. 149, 301–309.

Chang, J.J., Lin, J.-J., Ho, C.-Y., Chin, W.-C., Huang, C.C., 2010. Establishment of rumen-mimic bacterial consortia: a functional union for bio-hydrogen production from cellulosic bioresource. Int. J. Hydrogen Energy 35, 13399–13406.

Christopherson, M.R., Dawson, J.A., Stevenson, D.M., Cunningham, A.C., Bramhacharya, S., Weimer, P.J., Kendziorski, C., Suen, G., 2014. Unique aspects of fiber degradation by the ruminal ethanologen Ruminococcus albus 7 revealed by physiological and transcriptomic analysis. BMC Genomics 15, 1066.

Datta, R., 1981. Acidogenic fermentation of corn stover. Biotechnol. Bioeng. 23, 61–77.

De Souza, A.C., Rietkerk, T., Selin, C.G.M., Lankhorst, P.P., 2013. A robust and universal NMR method for the compositional analysis of polysaccharides. Carbohydr. Polym. 95, 657–663.

Eichorst, S.A., Joshua, C., Sathitsuksanoh, N., Singh, S., Simmons, B.A., Singer, S.W., 2014. Substrate-specific development of thermophilic bacterial consortia using chemically pretreated switchgrass. Appl. Environ. Microbiol. 80, 7423–7432.

Feng, Y., Yu, Y., Wang, X., Qu, Y., Li, D., He, W., Kim, B.H., 2011. Degradation of raw corn stover powder (RCSP) by an enriched microbial consortium and its community structure. Bioresour. Technol. 102, 742–747.

Gao, Z.M., Xu, X., Ruan, L.W., 2014. Enrichment and characterization of an anaerobic cellulolytic microbial consortium SQD-1.1 from mangrove soil. Appl. Microbiol. Biotechnol. 98, 465–474.

Guo, P., Zhu, W., Wang, H., Lü, Y., Wang, X., Zheng, D., Cui, Z., 2010. Functional characteristics and diversity of a novel lignocelluloses degrading composite microbial system with high xylanase activity. J. Microbiol. Biotechnol. 20, 254–264.

Hendriks, A.T.W.M., Zeeman, G., 2009. Pretreatments to enhance the digestibility of lignocellulosic biomass. Bioresour. Technol. 100, 10–18.

Himmel, M.E., Xu, Q., Luo, Y., Ding, S.-Y., Lamed, R., Bayer, E.A., 2010. Microbial enzyme systems for biomass conversion: emerging paradigms. Biofuels 1, 323–341.

Hollister, E.B., Hammett, A.M., Holtzapple, M.T., Gentry, T.J., Wilkinson, H.H., 2011. Microbial community composition and dynamics in a semi-industrial-scale facility operating under the MixAlco™ bioconversion platform. J. Appl. Microbiol. 110, 587–596.

Hui, W., Jiajia, L., Yucai, L., Peng, G., Xiaofen, W., Kazuhiro, M., Zongjun, C., 2013. Bioconversion of un-pretreated lignocellulosic materials by a microbial consortium XDC-2. Bioresour. Technol. 136, 481–487.

Hu, Z.H., Yu, H.Q., 2005. Application of rumen microorganisms for enhanced anaerobic fermentation of corn stover. Process Biochem. 40, 2371–2377.

Jiménez, D.J., Dini-Andreote, F., van Elsas, J.D., 2014. Metataxonomic profiling and prediction of functional behaviour of wheat straw degrading microbial consortia. Biotechnol. Biofuels 7, 92.

Kleerebezem, R., van Loosdrecht, M.C., 2007. Mixed culture biotechnology for bioenergy production. Curr. Opin. Biotechnol. 18, 207–212.

Kong, Y., Teather, R., Forster, R., 2010. Composition, spatial distribution, and diversity of the bacterial communities in the rumen of cows fed different forages. FEMS Microbiol. Ecol. 74, 612–622.

Krause, D.O., Denman, S.E., Mackie, R.I., Morrison, M., Rae, A.L., Attwood, G.T., McSweeney, C.S., 2003. Opportunities to improve fiber degradation in the rumen: microbiology, ecology, and genomics. FEMS Microbiol. Rev. 27, 663–693.

Lamed, R., Setter, E., Kenig, R., Bayer, E.A., 1983. Cellulosome: a discrete cell surface organelle of Clostridium thermocellum which exhibits separate antigenic, cellulose-binding and various cellulolytic activities. Biotechnol. Bioeng. Symp. 13, 163–181.

Lynd, L.R., Weimer, P.J., van Zyl, W.H., Pretorius, I.S., 2002. Microbial cellulose utilization: fundamentals and biotechnology. Microbiol. Mol. Biol. Rev. 66, 506–577.

McMurdie, P.J., Holmes, S., 2013. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8, e61217.

Monlau, F., Barakat, A., Steyer, J.P., Carrere, H., 2012. Comparison of seven types of thermo-chemical pretreatments on the structural features and anaerobic digestion of sunflower stalks. Bioresour. Technol. 120, 241–247.

Nishiyama, T., Ueki, A., Kaku, N., Watanabe, K., Ueki, K., 2009. Bacteroides graminisolvens sp. nov., a xylanolytic anaerobe isolated from a methanogenic reactor treating cattle waste. Int. J. Syst. Evol. Microbiol. 59, 1901–1907.

Nozière, P., Ortigues-Marty, I., Loncke, C., Sauvant, D., 2010. Carbohydrate quantitative digestion and absorption in ruminants: from feed starch and fibre to nutrients available for tissues. Animal 4, 1057–1074.

Pajual T., O'Donohue, M.J., 2013. Building Tomorrow's Biorefineries. Annexe D1.1. Biocore Report. <http://www.biocore-europe.org/file/D1_1> Availability of lignocellulosic biomass types of interest in the study regions.pdf.

Pfennig, N., Trüper, H.G., 1992. The family Chromatiaceae. In: Balows, A., Trüper, H.G., Dworkin, M., Harder, W., Schleifer, K.-H. (Eds.), The Prokaryotes. Springer, New York, pp. 3200–3221.

Prins, R.A., van Vugt, F., Hungate, R.E., van Vorstenbosch, C.J.A.H.V., 1972. A comparison of strains of Eubacterium cellulosolvens from the rumen. Antonie van Leeuwenhoek 38, 153–161.

Ramsak, A., Peterka, M., Tajima, K., Martin, J.C., Wood, J., Johnston, M.E.A., Aminov, R.I., Avgustin, G., 2000. Unravelling the genetic diversity of ruminal bacteria belonging to the CFB phylum. FEMS Microbiol. Ecol. 33, 69–79.

Reddy, A., Allgaier, M., Singer, S., Hazen, T., Simmons, B., Hugenholtz, P., VanderGheynst, J., 2011. Bioenergy feedstock-specific enrichment of microbial populations during high-solids thermophilic deconstruction. Biotechnol. Bioeng. 108, 2088–2098.

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.J., Weber, C.F., 2009. Introducing Mothur: opensource, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537–7541.

Ten Have, R., Teunissen, P.J., 2001. Oxidative mechanisms involved in lignin degradation by white-rot fungi. Chem. Rev. 101, 3397–3413.

Trifonova, R., Postma, J., Ketelaars, J.J.M.H., van Elsas, J.D., 2008. Thermally treated grass fibers as colonizable substrate for beneficial bacterial inoculum. Microb. Ecol. 56, 561–571.

Yan, L., Gao, Y., Wang, Y., Liu, Q., Sun, Z., Fu, B., Wen, X., Cui, Z., Wang, W., 2012. Diversity of a mesophilic lignocellulolytic microbial consortium which is useful for enhancement of biogas production. Bioresour. Technol. 111, 49–54.

III. Conclusion du chapitre

La sélection et la stabilisation d'une communauté bactérienne active sur lignocellulose en conditions anaérobie est possible à partir d'un inoculum de rumen bovin. La communauté obtenue, stabilisée en dix cycles d'enrichissement, a été appelée RWS. Elle présente une activité xylanase élevée, majoritairement non libre, et dégrade simultanément la cellulose et l'hémicellulose du substrat en proportions égales. Elle est capable de dégrader 55% en poids sec du substrat utilisé (de la paille de blé) en produisant principalement des acides gras volatils, ce qui en fait un bon candidat pour une utilisation en plateforme carboxylate, en conditions anaérobies, non stériles et en absence de prétraitement.

Après dix cycles de culture, la diversité de RWS apparait stable d'un cycle de culture à l'autre. Sa diversité est très réduite par rapport à l'inoculum initial, mais est composée principalement de *Bacteroidetes* et *Firmicutes*, tous deux des représentants classiques de la flore ruminal bovine. RWS apparait donc comme un enrichissement de bactéries ruminales adaptées aux conditions de culture en fermenteur et dont les capacités de dégradation sont augmentées par rapport à l'inoculum initial.

Le suivi de la dégradation du substrat, de la production d'acides gras volatils et des activités enzymatiques au cours d'un cycle de dégradation par RWS montre que ces différents paramètres n'évoluent pas de façon linéaire au cours du temps. Afin de mieux comprendre comment fonctionne la communauté bactérienne, et notamment si elle évolue en cours de dégradation, en composition ou en activité, il est intéressant d'étudier sa dynamique au cours d'un de ces cycles. Par ailleurs, l'utilisation de substrats prétraités peut permettre à la fois d'augmenter les capacités de dégradation de RWS, mais aussi d'identifier les populations bactériennes sensibles à des variations du substrat utilisé.

CHAPITRE V

ETUDE DE L'IMPACT DU PRETRAITEMENT DU SUBSTRAT SUR LA DYNAMIQUE DES COMMUNAUTES BACTERIENNES AU COURS DE LA DEGRADATION

CHAPITRE V : ETUDE DE L'IMPACT DU PRETRAITEMENT DU SUBSTRAT SUR LA DYNAMIQUE DES COMMUNAUTES BACTERIENNES AU COURS DE LA DEGRADATION

I. Introduction du chapitre

La dégradation de la lignocellulose est un processus complexe, faisant intervenir de nombreuses activités enzymatiques complémentaire pour l'attaque des différents composants du substrat. Cette nécessité d'une grande diversité enzymatique explique les difficultés pour dégrader la lignocellulose à l'aide de cocktails enzymatiques ou de cultures pures. Mais dans le cas de l'utilisation d'une communauté microbienne, les populations qui la composent peuvent chacune jouer des rôles différents, à des moments différents. Cependant, dans la plupart des travaux, la diversité bactérienne n'est étudiée qu'en un seul point, généralement au temps final de la dégradation.

Le suivi des dynamiques de la diversité (par séquençage de l'ADNr 16S) et de l'activité (par séquençage de l'ARNr 16S) bactérienne permet d'identifier les populations associées aux différentes phases de la dégradation de lignocellulose en fermenteur, et donc les populations responsables de la dégradation observée. Dans ce contexte, l'utilisation des transcrits de l'ADNr 16S (l'ARNr 16S), permettant de s'intéresser aux populations actives, plutôt qu' au gène codant pour l'ADNr 16S, par ailleurs moins variable, permet une meilleure caractérisation des liens entre dégradation et populations bactériennes actives.

Enfin, l'utilisation de substrats modifiés par prétraitement permet de caractériser à la fois leur impact sur les capacités de dégradation (partie qui est l'objet d'une étude détaillée par A. Lazuka, et fera l'objet d'une publication don je suis co-auteur) mais aussi sur la communauté bactérienne et son activité. Jusqu'à présent, si certaines études se sont intéressées à l'effet du prétraitement d'un substrat sur les communautés pendant l'enrichissement (Reddy et al., 2012; Eichorst et al., 2014), aucune n'a caractérisé l'impact du prétraitement sur la dynamique d'une communauté pendant un cycle de dégradation.

Dans ce chapitre, l'effet du prétraitement du substrat, modifié par des méthodes innovants par voie sèche combinant le broyage et l'imprégnation à la soude, sur les dynamiques d'activité bactériennes sera étudié. Notre objectif est d'identifier les populations bactériennes par séquençage de l'ARNr 16S dont l'activité varie en fonction des phases de dégradation ou du prétraitement appliqué au substrat.

II. Effect of lignocellulosic substrate pretreatment on lignocellulolytic community dynamics

Ce chapitre sera soumis à « Environmental Microbiology » sous le titre :

Diversity of metabolically active bacteria associated with lignocellulose degradation by a cow rumen-derived consortium: dynamic response to substrate pretreatment.

Lucas Auer[1,2,3 *], Adèle Lazuka[1,2,3 *], Michael O'Donohue[1,2,3], Guillermina Hernandez-Raquet[1,2,3*]

[*] These authors contributed equally to this work.
[1)] Université de Toulouse, INSA, UPS, LISBP, 135 Avenue de Rangueil, F-31077 Toulouse Cedex 4, France
[2)] INRA, UMR792 Ingénierie des Systèmes Biologiques et des Procédés, F-31400 Toulouse, France
[3)] CNRS, UMR5504, F-31400 Toulouse, France

II.1. Abstract

Pretreatment of lignocellulosic substrates is an unavoidable step in all biomass bioconversion processes, including the carboxylates platform. While the effects of pretreatment on substrate structure and composition have been extensively studied, the response of metabolically active bacteria to substrate modifications induced by pretreatment is still not well understood. This study assessed the functional community response to fermentation of mechanical and dry-chemo-mechanical pretreated wheat straw. Two milling conditions (2 mm and 100 µm) associated or not to dry-NaOH impregnation were the applied biomass pretreatments. These pretreated substrates were anaerobically fermented by a rumen-derived bacterial consortium RWS in controlled bioreactors. The diversity of the whole community and the metabolically active bacteria were monitored by sequencing, respectively, the 16S rRNA gene (DNA) and its transcripts (RNA) through the lignocellulose degradation processes. Irrespective of the pretreatment used, diversity of the metabolically active community showed a cyclic behavior, displaying a similar composition at the initial and final times of incubation while a distinct community makeup was observed when the consortium displayed the maximum lignocellulose degradation rates. For all pretreatments, the 16S rDNA data showed a strong dominance of few OTUs related to *Bacteroides, Prevotella* and *Rikenellaceae*. In contrast, OTU predominance was weaker when considering the diversity of the active community (16S rRNA data). Nevertheless, the diversity's cyclic behavior was maintained, with a clear separation in three phases corresponding to the different

lignocellulose degradation phases. During the initial lignocellulose degradation phase, a highly diverse community was observed with a strong transcription of *Bacilli* related OTUs. When the maximum lignocellulose degradation rates were measured, diversity was characterized by a strong transcription of *Bacteroides* and *Clostridia* related OTUs. At the end of the incubation, when lignocellulose degradation rate decreased, diversity showed a shift to the initial community composition. This study presents the first report on the dynamics response of metabolically active bacteria through bioconversion of pretreated lignocellulosic substrates.

II.2. Introduction

Lignocellulose (LC) is the main component of plant biomass and represents the most promising renewable carbon source on Earth (Chandel and Singh, 2011). In the last decade, non-food LC has been receiving much attention to reduce fossil carbon dependence and valorize a resource so far considered as a waste. Different strategies are currently considered for LC valorization. The enzymatic deconstruction of LC into simple sugar (sugar platform) followed by pure-culture fermentation is the most studied approach for bio-ethanol production. The bioconversion of LC using mixed microbial communities to produce biogas (bio-H_2, bio-CH_4) or carboxylates (carboxylate platform) has also been proposed. However, irrespective of the target product, LC bioconversion is still challenging. Indeed, LC consists of three main polymers, cellulose, hemicelluloses and lignin, which are associated to each other in a complex structure conferring biomass a strong resistance to degradation (Mosier et al., 2005). Thus, all LC bioconversion strategies utilize pretreatment methods to reduce biomass recalcitrance and improve transformation yields. The objective of pretreatments is to modify the crystalline structure of LC, reduce its lignin content, and increase its accessible surface and porosity. The impact of pretreatment on substrate structure and composition and its consequences on LC digestibility has been extensively studied (Mosier et al., 2005; Hendriks and Zeeman, 2009; Guo et al., 2011; Monlau et al., 2012). However, except some evidences of inhibition of ethanologenic yeast induced by furan and lignin-derived compounds produced by pretreatment, little is known about the microbial community's response to substrate modifications induced by biomass pretreatment. Previous studies comparing the diversity of microbial communities growing in various lignocellulose feedstocks displaying different physicochemical features (eucalyptus, switchgrass) showed that similar microbial communities were selected in each particular substrate, irrespective of

the initial diversity of the inoculum (Simmons et al; 2014). Other studies revealed significant differences between communities selected in different LC biomass (switchgrass and corn stover), demonstrating that some OTUs were exclusively associated to a specific substrate (Reedy et al., 2011). Similarly, in microbial enrichments using pretreated switchgrass by ammonia fiber expansion or ionic liquids, the diversity of the enriched communities was specific to each substrate (Eichorst et al., 2014). These studies suggest that the specific composition of LC feedstock determines the diversity of the selected microbial community. However, most of these studies compared the diversity of microbial communities enriched in different feedstocks at the end of the incubation, when lignocellulose transformation was already stopped or it displayed a low degradation rate. In the other hand, it is known that metabolically active bacteria evolve throughout the bioconversion processes, particularly when degrading complex substrates. It would be expected that the diversity of metabolically active bacteria would adapt in response to substrate and environmental modifications induced by its own activity (Snajdr et al., 2011). Based on these observations it is possible to ask: Which is the diversity of species metabolically active on lignocellulosic biomass? Are they common irrespective the substrate pretreatment applied? And, which are the metabolically active species acting when the strongest LC degradation occurs? Investigating the microbial communities metabolically active when the community displays the highest LC transformation rates, on diverse pretreated substrates, can provide insights on the more efficient LC deconstructing bacteria to favor in LC biorefinery industrial scale processes.

Mechanical (generally milling) and chemical pretreatment (especially alkaline) are two well established pretreatment approaches. Particle size reduction induced by milling increases the accessible surface (Hendriks and Zeeman, 2009). In chemical alkaline pretreatments, sodium hydroxide strongly modify lignin and induce cellulose swelling, improving biomass porosity and accessibility (Zhao et al., 2008). Simultaneous mechanical and chemical treatments, and particularly the recently developed dry-chemo-mechanical pretreatments have showed a strong enhancement of biomass digestibility, representing an eco-friendly and economically sustainable pretreatment (Barakat et al., 2014). Milling and dry-chemo mechanical pretreatments modify the accessible surface of biomass, its crystallinity and macro porosity increasing significantly the biomass digestibility by a rumen-derived microbial consortium compared to raw biomass (Lazuka et al., 2016). However, while the impact of pretreatment on the microbial bioconversion potential has been described, its effect

on the diversity of metabolically active bacteria along the degradation processes has not yet been investigated.

The aim of this study was to assess the changes on metabolically active bacteria induced by mechanical and dry-chemo-mechanical pretreatment of wheat straw. The diversity of functional bacterial communities was monitored by sequencing the transcripts of V3-V4 16S rRNA gene region throughout the lignocellulose degradation process by a rumen-derived consortium. The dynamic diversity analysis of metabolically active bacteria growing on different pretreated substrates showed that despite diversity varied in function of the pretreated biomass feature's, the diversity profiles at the peak of LC degradation were similar. These results provide evidence of successive adaptation of the community to lignocellulosic substrate changes occurring during cultivation.

II.3. Methods

II.3.1 Lignocellulosic pretreated substrates

Wheat straw of the Koreli variety was milled to 2 mm using a knife mill (Retsch SM 100, Germany). It was considered as the reference material and is thereafter named pretreatment A. Dry sodium hydroxide impregnation combined or not to milling enabled to produce pretreated straws B: 2mm, NaOH; C: 100µm and D: 100µm, NaOH, which were obtained as previously described (Barakat et al., 2014, Lazuka et al., 2016). The physicochemical characterization of these substrates was previously described (Lazuka et al., 2016); their main substrate features are listed in Table 1.

II.3.2 Lignocellulose degradation reactors

Lignocellulose degradation was assessed in anaerobic batch reactors (2L BIOSTAT® A+, Sartorius, Germany) using a mineral medium (Lazuka et al. 2015) containing the different pretreated wheat straws (A-D) as sole carbon source ($20g.L^{-1}$). Reactors were inoculated (10% v/v) with a cow-rumen derived consortium RWS displaying a good wheat straw degrading ability (Lazuka et al., 2015). Bioreactors were operated with controlled pH (6.15) under mesophilic conditions (35°C) in order to favor carboxylate production, as detailed by Lazuka et al. (2015). The lignocellulose degradation rates throughout incubation of different pretreated substrates were estimated base on lignocellulose concentration loss determined as volatile solids. Volatile solids (VS) were estimated from the difference between

total solids (TS) and mineral solids (MF). TS were measured using 10 mL samples, centrifuged (7,197 x g; 10 min), rinsed twice with distilled water and dried 24h at 105°C. The MF was estimated by mineralization of the samples at 500°C for 2h. For each treatment, average degradation was obtained after polynomial regression applied to degradation values of the two biological replicates.

II.3.3 Nucleic acids isolation

Samples (1.5mL) were taken at regular intervals and immediately centrifuged (13,000 g, 5 min, and 4°C). After removing the supernatant, the pellet was snap frozen in liquid nitrogen and stored at -80°C until nucleic acid extraction. Total DNA and RNA were simultaneously extracted using a PowerMicrobiome RNA Isolation kit (MoBio Laboratories Inc. Carlsbad) following the manufacturer's instructions but omitting the final DNAse steps. Cell lysis was carried out with a Fast Prep (MP Biomedicals) using 2 cycles of 30s at $4ms^{-1}$. Subsequently, DNA and RNA were separated and purified using an AllPrep DNA/RNA Mini Kit (Qiagen) according to the manufacturer's instructions. DNA and RNA integrity and purity were checked by agarose gel (1%) electrophoresis and concentrations were measured by NanoDrop 1000 spectrophotometer (Thermo Scientific). Residual DNA eventually present in RNA samples was removed with TURBO DNA-free™ kit (Ambion, Life Technologies) according to the manufacturer's instructions. RNA quality was checked on an RNA Pico 6000 Chip Kit, using a Bioanalyzer 2100 (Agilent Technologies). Reverse transcription (RT) was performed in RNA extracts using M-MLV Reverse Transcriptase (Promega) and random hexamers (Roche), following the manufacturer's instructions. cDNA was stored at -80°C until amplification.

II.3.4 DNA and cDNA amplification and sequencing

V3-V4 16S rRNA gene region was amplified by PCR using the modified bacterial primers 343F and 784R:

343F=5'-CTT-TCC-CTA-CAC-GAC-GCT-CTT-CCG-ATC-TAC-GGR-AGG-CAG-CAG-3'

784R=5'-GGA-GTT-CAG-ACG-TGT-GCT-CTT-CCG-ATC-TTA-CCA-GGG-TAT-CTA-ATC-CT-3'

PCR1 was performed in 50µl reaction mixture containing 1X PCR buffer, 2.5U MTP Taq DNA Polymerase (Sigma), 0.2µM of each dNTP, 0.5µM of each primer and using 2ng of DNA samples, or 1µL of a 20 fold dilution of cDNA. After 30 amplification cycles at 94°C, 65°C, 70°C, one minute each step, amplicons were purified using magnetic beads and

quantified by NanoDrop 1000 spectrophotometer. A second PCR was performed to add sequencing adapters and a unique index for each sample using the primers FP2= 5'-AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC-3' and RP2= 5'-CAA GCA GAA GAC GGC ATA CGA GAT-index-GTG ACT GGA GTT CAG ACG TGT-3'. PCR was performed in 50µl reaction mixture containing 1X PCR buffer, 2.5U MTP Taq DNA Polymerase (Sigma), 0.2µM of each dNTP, 0.5µM of each primer and 15ng of previously amplified DNA. After 12 amplification cycles (94°C, 65°C, 72°C, one minute each step), amplicons were purified using magnetic beads and quantified by Nanodrop 1000 spectrophotometer. Amplicon quality was then checked with High Sensivity DNA Analysis Kits (Agilent) and a BioAnalyzer 2100. An equimolar pool was then prepared and loaded on a MiSeq Illumina cartridge, using reagent kit v3 (paired 300bp reads). Sequencing was performed at the GenoToul Genomics and Transcriptomics facility (GeT, Auzeville, France) using a MiSeq® Illumina®.

II.3.5 Data processing

Data were demultiplexed by the GeT platform, which also joined the pair-end reads with Flash v1.2.6 (Magoč and Salzberg, 2011), with an minimum overlap of 110bp a maximum ratio of 0.1 mismatches. Associated with the MiSeq® intern filters, these parameters allowed to perform a good quality-filter. All the fastq files were then merged into a unique fasta file, treated with Mothur v1.33.1 (Schloss et al., 2009) following the SRF1 procedure described in Auer et al. (2016). Briefly, sequences presenting primer mismatches or an unexpected length were discharged before proceed to dereplication step. Singleton reads and low-quality reads were discarded. Sequences were then aligned against SILVA database and only those aligning to the expected V3-V4 region were further preclustered with d=5. After a chimera detection by Uchime (Edgar et al., 2011) using a self-reference, sequences were clustered using Mothur's average-distance algorithm with a 3% threshold. Rare OTUs containing less than 100 sequences across overall dataset, equivalent to less than 0.001% (under the 0.005% threshold recommended by Bokulich et al. (2013)) were removed. OTUs were taxonomically affiliated based on the consensus assignation of its constitutive sequences, determined with Mothur's implementation of RDP Classifier and the fusion of LTP SSU database version 115 (Yarza et al., 2008) and DictDB (Mikaelyan et al., 2015). OTU tables were constructed with Mothur, and OTU fasta files were generated using the most

abundant sequence of each OTU. Phylogenetic trees were generated using ClustalO (Sievers et al., 2014) and raxmlHPC (Stamatakis, 2014).

II.3.6 Diversity analysis

OTU abundance tables, taxonomy files and phylogenetic trees were manually imported into R v3.2.3 using the Phyloseq package version 1.14.0 (McMurdie and Holmes, 2013), following the procedure described on http://joey711.github.io/phyloseq/import-data. Richness, diversity indexes (Shannon and Simpson), weighted-Unifrac distances between samples and PCoA ordinations (were calculated and plotted using Phyloseq. Sparse-PLS discriminant analyses (sPLS-DA) were performed using mixOmics v5.2.0 package (Cao et al., 2009). Statistical differences between variables were tested using ANOVA with R basic functions anova and lm. PERMANOVA analyses were performed using adonis function of the vegan package v2.3.5.

II.4. Results

II.4.1 Degradation of pretreated wheat straw substrates by RWS

Duplicate biological reactors were established with dry-mechanically (A: 2 mm or C: 100 µm size) or dry-chemo-mechanically (B: 2mm, NaOH and D: 100µm, NaOH) pretreated wheat straw as sole carbon source. To assess the lignocellulose degradation efficiency of each pretreated substrate, bioreactors were inoculated with a rumen-derived consortium RWS. Lignocellulose degradation, determined as the loss of volatile solids along the incubation time in two biological replicates foe each substrate, was significantly higher for alkali pretreated substrates B (2mm, NaOH) and D (100µm, NaOH), reaching a maximal LC degradation of, respectively, 64±5% and 49±7% (Figure 1). In contrast, mechanically pretreated substrates A (2mm) and C (100µm) displayed lower LC transformation levels of 56±3% and 33±1%, respectively, at the end of the incubation. Particularly, dry-alkali pretreated substrate displayed the fastest degradations rates reaching 2.25g.day^{-1} and 2.21g.day^{-1}, respectively (Figure 1). Irrespective of the pretreated substrate, lignocellulose degradation rate increased at the begging on the incubation to reach a maximum value; thereafter it decreased again. The maximal degradation rates were observed earlier for dry-alkali pretreated substrates, around in average 3-4 days of incubation, compared to substrates only treated by comminution, which displayed a maximal degradation rate after day 4-5 in average.

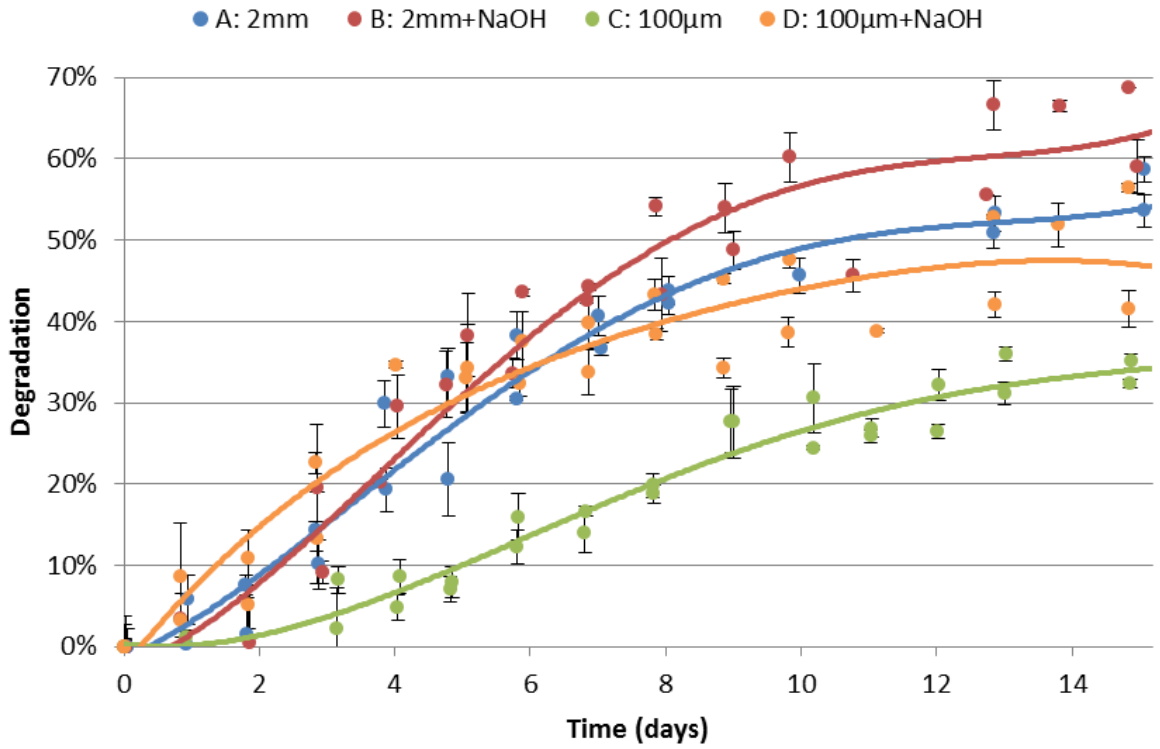


Figure 1. Degradation percentage during each replicated incubation according to pretreatments A (2mm), B (2mm-NaOH), C (100µm) and D (100µm-NaOH). Experimental points with their standard deviation are indicated in black, lines correspond to average degradation between two biological replicates, calculated by polynomial regressions.

II.4.2 Microbial community structure on different pretreated substrates

To get insight into the community structure corresponding to each pretreatment and lignocellulose degradation rate, sequencing was performed on samples taken regularly throughout the incubation period. Sequencing of the V3-V4 16S rRNA gene region (DNA) and its transcripts (RNA) was used to characterize the whole bacterial diversity (DNA) and the diversity of metabolically active bacteria (RNA) associated with the transformation of each pretreated substrate.

Overall 15M sequences were generated by MiSeq Illumina sequencing, corresponding to an average of 60,000 sequences per sample. Length and primer mismatch filtering discharged 5% of sequences. Singleton reads were removed after dereplication, discharging 26% of the initial sequences. Subsequent alignment filter and chimera detection resulted in 10.6M high-quality sequences which were clustered into 4,065 OTUs. OTUs containing less than 100 sequences were removed resulting in a final dataset of 10,58M high-quality

sequences containing 455 OTUs. Rarefaction curves generated from the sequencing data of each sample indicated a sufficient coverage of the microbial communities of each sample (Supplementary data, Figure S1).

The kinetics of richness observed with the different substrates showed a global decrease during the first days of incubation, thereafter the richness level increase again. Similarly, the kinetics of the microbial community's diversity, based on Simpson and Shannon index (estimated based on OTUs constructed by clustering at 97% homology) displayed a similar behavior with a decrease in the first days of incubation. However, such decrease in richness and diversity were more pronounced on dry-chemical pretreated substrate (B and D) compared to mechanically treated substrates A and C (Figure 2). The community diversity estimated for substrates A and C displayed average values of Simpson=0.85 and Shannon=2.6. In contrast, the impact on the dynamics of richness and diversity was stronger on pretreated substrates B and D, displaying a sharped decrease (Shannon<2 and Simpson<0.7) during the first days of incubation followed to a return to their initial levels.

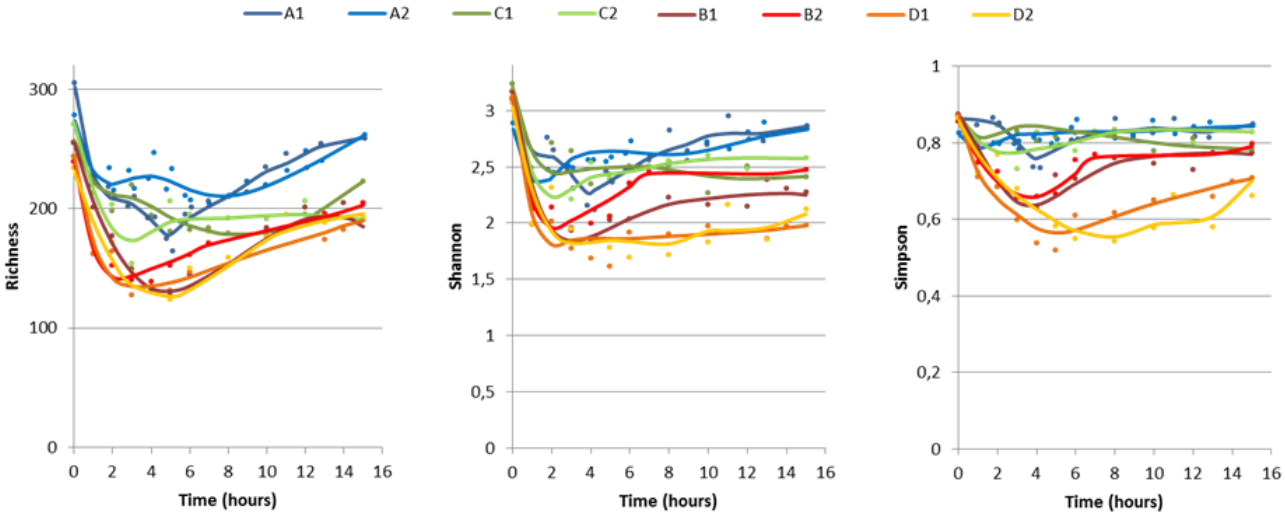


Figure 2. Richness, Shannon and Simpson indexes evolution during fermentation, according to pretreatments A (2mm), B (2mm-NaOH), C (100µm) and D (100µm-NaOH). Experimental points are indicated with dots, lines correspond to polynomial regression.

PCA ordination of the average community composition observed with each treatment using DNA and RNA data revealed differences in microbial community makeup associated with each pretreated substrates (Figure 3). PCA showed clear separation between microbial communities, where the dry-chemically pretreated substrates were enriched in *Bacteroidia*

whereas non-chemically treatments displayed more *Bacilli* and *Clostridia*. Moreover, treatments B and D showed variability between their respective replicates, due to *Spirochaetes* and *Fibrobacteres* content of one of the replicates.

Accordingly, sPLS-DA performed using OTUs of metabolically active bacteria (16S rRNA transcript data) and pretreatment as discriminant variable, revealed a shift in the metabolically active community within each pretreated substrate. sPLS-DA clearly discriminate chemically treated substrates (B and D) from those non-chemically treated (A and C) (Figure 4). Moreover, sPLS-DA analysis enabled to identify OTUs related with each pretreatment. Among the discriminating OTUs identified by sPLS-DA, 8 belonged to *Clostridia* and 6 were related to *Bacteroidia* (Table 1). Particularly, OTU_1, OTU_6, OTU_{10} and OTU_{15}, affiliated to, respectively, *Bacteroides*, *Clostridium*, *Enterobactereriaceae* and *Clostridiales*, were strongly associated with chemically-pretreated substrates while 2 mm

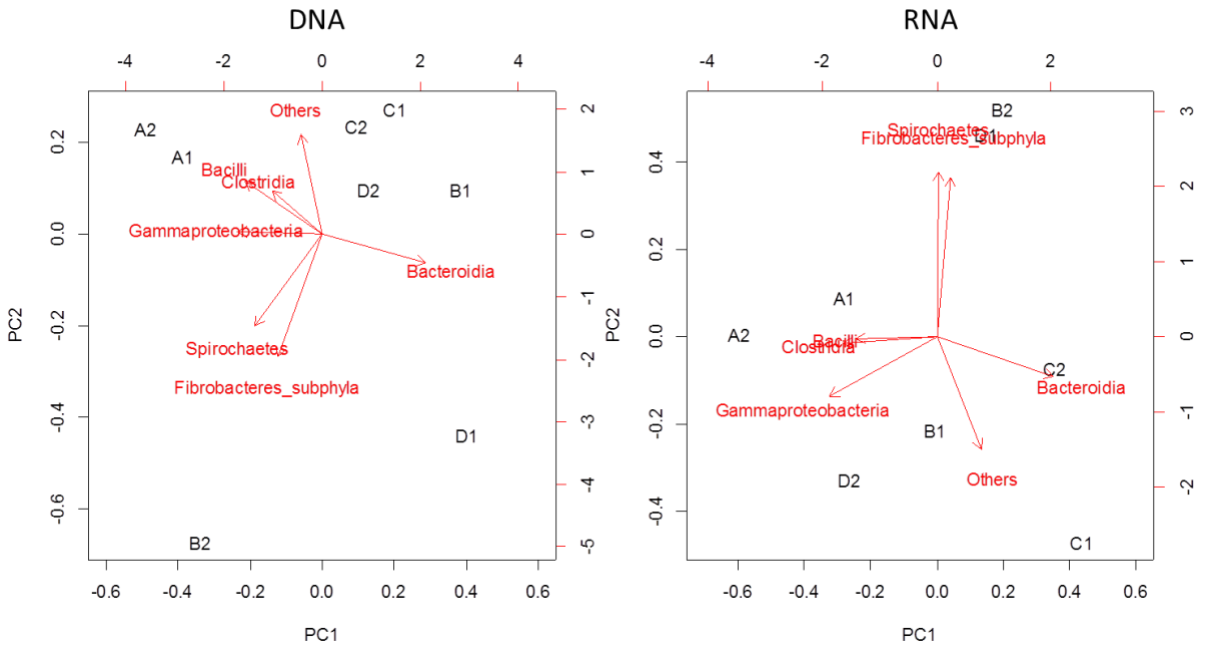


Figure 3. PCA plot performed with rDNA (left) and rRNA (right) data. Each point corresponds to the average profile of a replicate, at the microbial class level. Minor classes were regrouped in the "Others" category. Names correspond to pretreatments A (2mm), B (2mm-NaOH), C (100µm) and D (100µm-NaOH) and replicate number

Escherichia. Pretreated straw at 100 µm milling (C) was enriched with OTU_8 and OTU_{37}, belonging to, respectively, *Prevotella* and *Lachnospiraceae*. The statistical significance of differential expression of these OTUs was confirmed by Anova, with P-values all lower than 10^{-5} (Table 1).

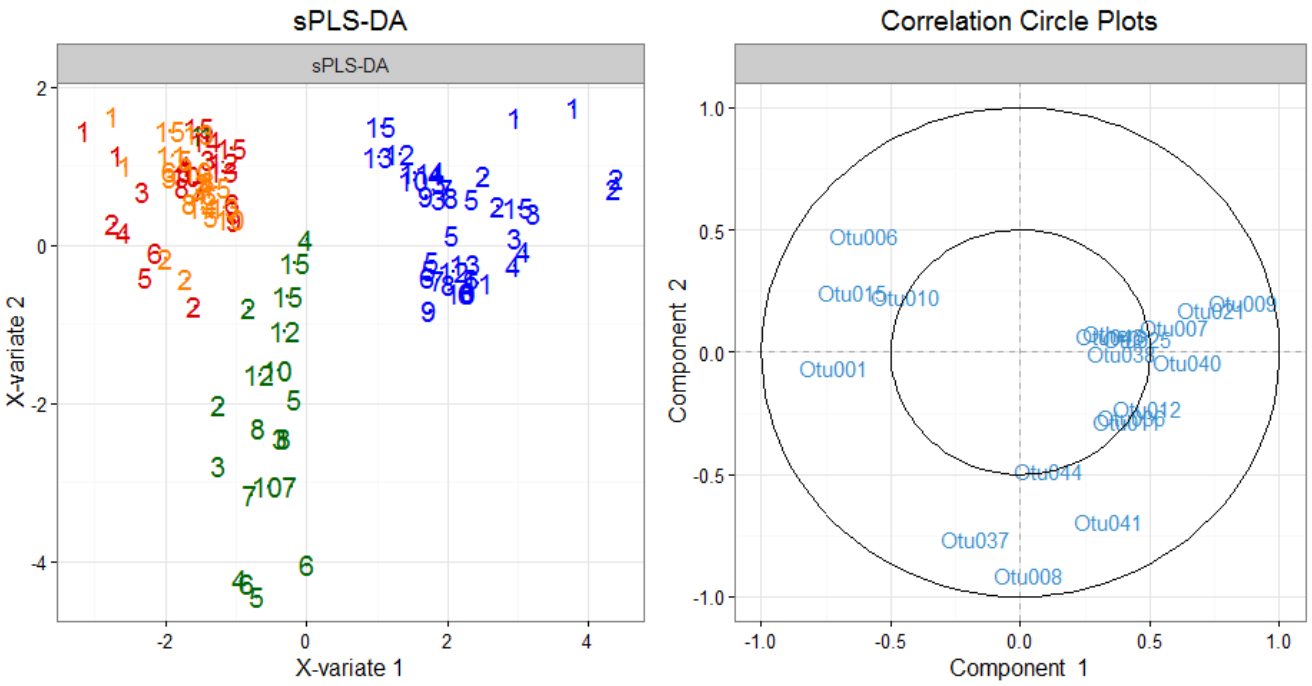


Figure 4. Sparse Partial Least Square Discriminant Analysis (sPLS-DA) analysis using RNA data and treatments as factor (left) and corresponding correlation circle (right). Colors correspond to treatments: 2mm (A) in blue, 2mmNaOH (B) in red, 100µm (C) in green and 100µmNaOH (D) in orange.

Table 1. rRNA expression level (relative to total rRNA library) of sPLS-DA selected OTUs, according to pretreatments. Percentages correspond to average level for each pretreatment. Taxonomy at the Class level and at the best level available is given for each OTU.

			Mean rRNA expression				ANOVA	
OTU	Class	Taxonomy	A	B	C	D	P-value	
1	*Bacteroidia*	*Bacteroides*	11,7%	29,2%	25,4%	28,1%	2,02E-12	***
6	*Clostridia*	*Clostridium*	2,5%	10,5%	4,5%	10,8%	1,80E-13	***
7	*Bacilli*	*Enterococcus*	11,1%	1,7%	4,2%	0,9%	1,86E-06	***
8	*Bacteroidia*	*Prevotella*	3,8%	2,5%	10,8%	1,8%	4,98E-13	***
9	*γ-proteobacteria*	*Escherichia-Shigella*	11,1%	1,7%	4,2%	0,9%	2,20E-16	***
10	*γ-proteobacteria*	*uncl Enterobacteriaceae*	8,8%	0,4%	1,2%	1,2%	8,25E-07	***
11	*Bacteroidia*	*Prevotella*	0,9%	0,2%	0,6%	0,2%	8,57E-06	***
12	*Bacteroidia*	*Prevotella*	2,2%	1,2%	1,4%	0,8%	1,71E-05	***
15	*Clostridia*	*uncl Clostridiales*	0,8%	4,0%	1,2%	4,5%	2,50E-13	***
21	*Bacteroidia*	*Dysgonomonas*	0,8%	0,01%	0,2%	0,1%	5,60E-13	***
25	*Clostridia*	*uncl Lachnospiraceae*	0,6%	0,04%	0,1%	0,04%	3,21E-07	***
36	*Clostridia*	*uncl Lachnospiraceae*	0,5%	0,1%	0,5%	0,0%	1,42E-07	***
37	*Clostridia*	*uncl Lachnospiraceae*	0,4%	0,5%	1,0%	0,6%	6,09E-07	***
38	*Clostridia*	*Oscillibacter*	0,5%	0,2%	0,3%	0,4%	2,84E-05	***
40	*α-proteobacteria*	*Rhodobacter*	0,6%	0,1%	0,2%	0,1%	3,38E-10	***
41	*Clostridia*	*uncl Clostridia*	0,5%	0,03%	0,6%	0,03%	1,08E-08	***
44	*Clostridia*	*Clostridium*	0,1%	0,03%	0,2%	0,05%	2,02E-05	***
45	*Bacteroidia*	*Prevotella*	0,7%	0,03%	0,03%	0,04%	3,37E-05	***
		Others	9,6%	5,3%	6,4%	5,6%	7,33E-05	***

II.4.3 Community dynamics during LC degradation

Further insight into the diversity of metabolically active bacteria within the four pretreated substrates was obtained comparing the temporal dynamics of active bacteria and the lignocellulose degradation rate observed in each pretreatment. Comparing Figure 5 with Figure 1, it is possible to remark that changes on the active community's composition occurred simultaneously with the strongest lignocellulose degradation rates. Community dynamics were investigated using Principle Coordinate Analyses (PCoA) based on weighted-Unifrac distances (Figure 6). A cyclic behavior of active communities was observed for all pretreatments.

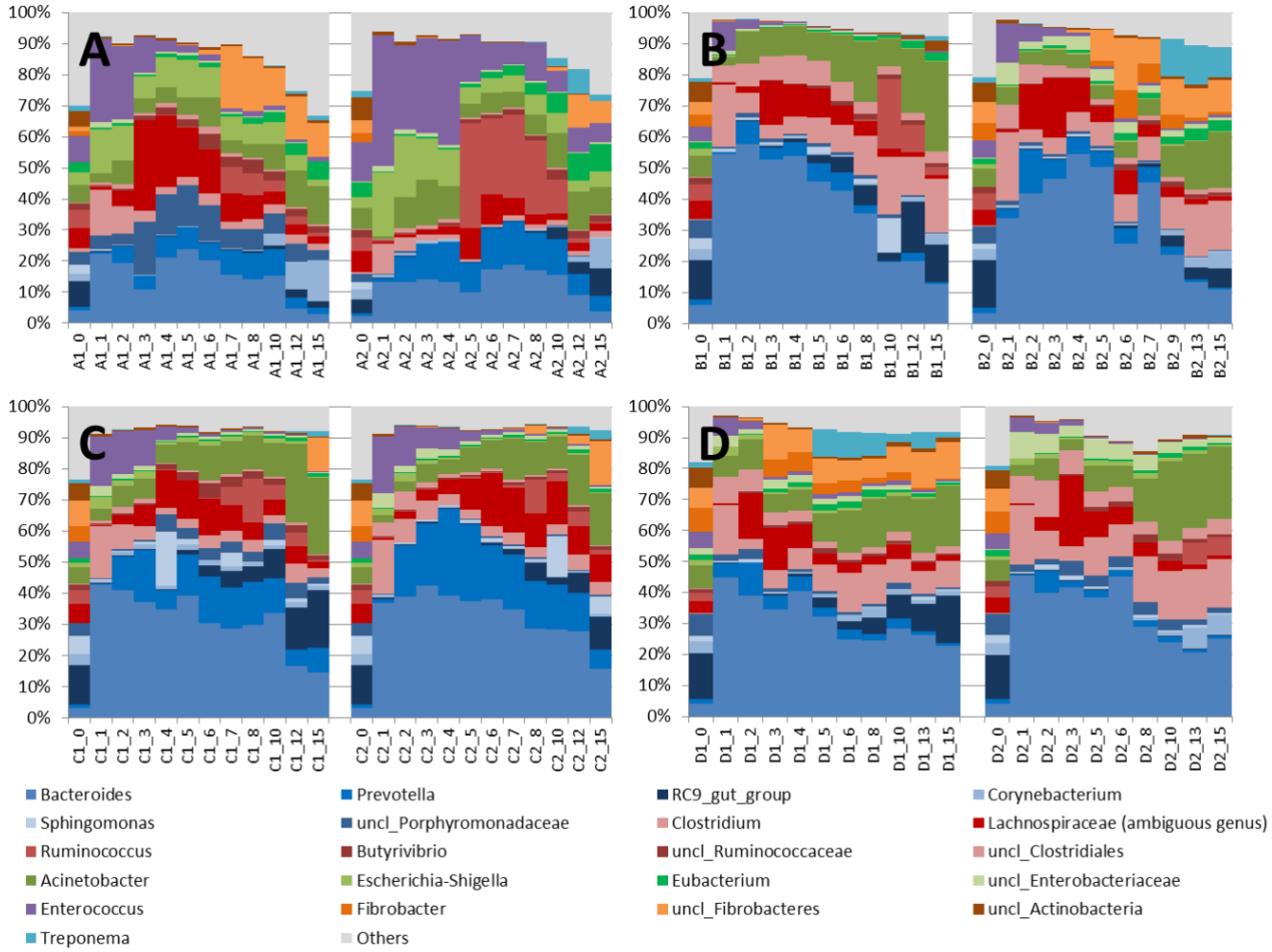


Figure 5. Diversity profiles at the genus level for each pretreatment and replicate. Colors are chosen according to Classes *Bacteroidia* (blue), *Clostridia* (red), *Gammaproteobacteria* (green), *Bacilli* (purple), *Fibrobacteres* (cyan), *Actinobacteria* (brown) and *Spirochaetes* (orange). Category Others contains minor genus that were less than 1% abundant.

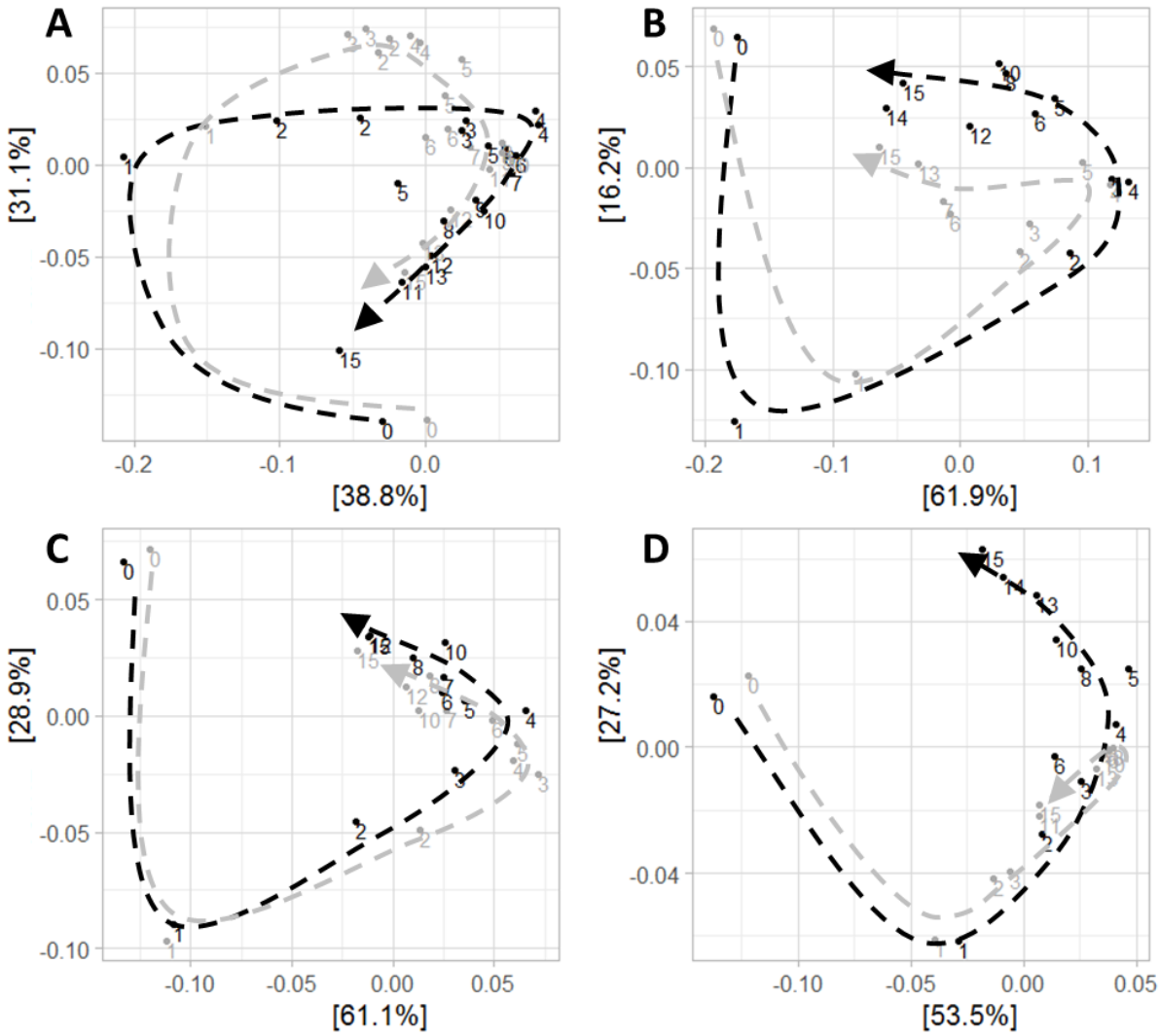


Figure 6. Principle Coordinate Analysis plots per treatment (based on weighted-Unifrac distances). A: 2mm, B: 2mmNaOH, C: 100µm, D: 100µmNaOH. Replicate 1 and 2 are respectively represented in black and gray. Percentages indicate the proportion of variability captured by each axes.

Three distinct phases during lignocellulose transformation were defined based on the measured lignocellulose degradation rates. Initial phase was characterized by a low degradation rate at the beginning of the incubation (t=0 excluded) ; second phase grouped the three points when the highest lignocellulose degradation rates were measured and last phase consisted in the last three time points, when degradation rates decreased and stabilized at the end of the incubation. Weighted-Unifrac distances between communities present at these three phases were calculated and PCoA analysis showed good separation between phases; the variance represented by two first components was of 90% (Supplementary). Moreover, PERMANOVA analysis explained 83% of weighted-Unifrac distances. The different phases (degrad_sep) explained 36% of the variance, the pretreatment method (Treatment) explained

30%, the interaction between these two factors explains 11% of this variance and replicates accounted for 6%. That means that variations observed between the active communities were mainly explained by the different phases observed on the kinetics of LC degradation but also by the y substrate pretreatment.

Metabolically active communities observed for each phase were compared by sPLS-DA using degradation phases as discriminant factor (Figure 7). Variable selection tuning resulted in 19 selected OTUs which are characteristic of each phase of degradation, including 5 *Bacteroidia* and 9 *Clostridia* members. LC degradation phases were well separated according to component 1 that separated the end phase of degradation, and component 2 for phases 1 and 2. Selected OTUs were checked for differential expression between such phases with an ANOVA analysis (Table 2). OTU_1 (*Bacteroides)* appeared specific of phase 1 and 2, whereas OTU_3, OTU_4 and OTU_{11} (respectively an ambiguous *Lachnospiraceae*, possibly a *Clostridium*, *Bacteroides* and *Prevotella*) were associated only with phase 2. OTU_6, OTU_7 and OTU_{30} (*Clostridium*, *Lactobacilles* and *Bacteroides*) were specific of initial points. The end degradation points were associated with *Rikenelllaceae* (OTU_2), *Acinetobacter* (OTU_5), *Eubacterium* (OTU_{18}), *Corynebacterium* (OTU_{19}), *Comamonadaceae* (OTU_{39}) and *Ruminococcaceae* (OTU_{49}).

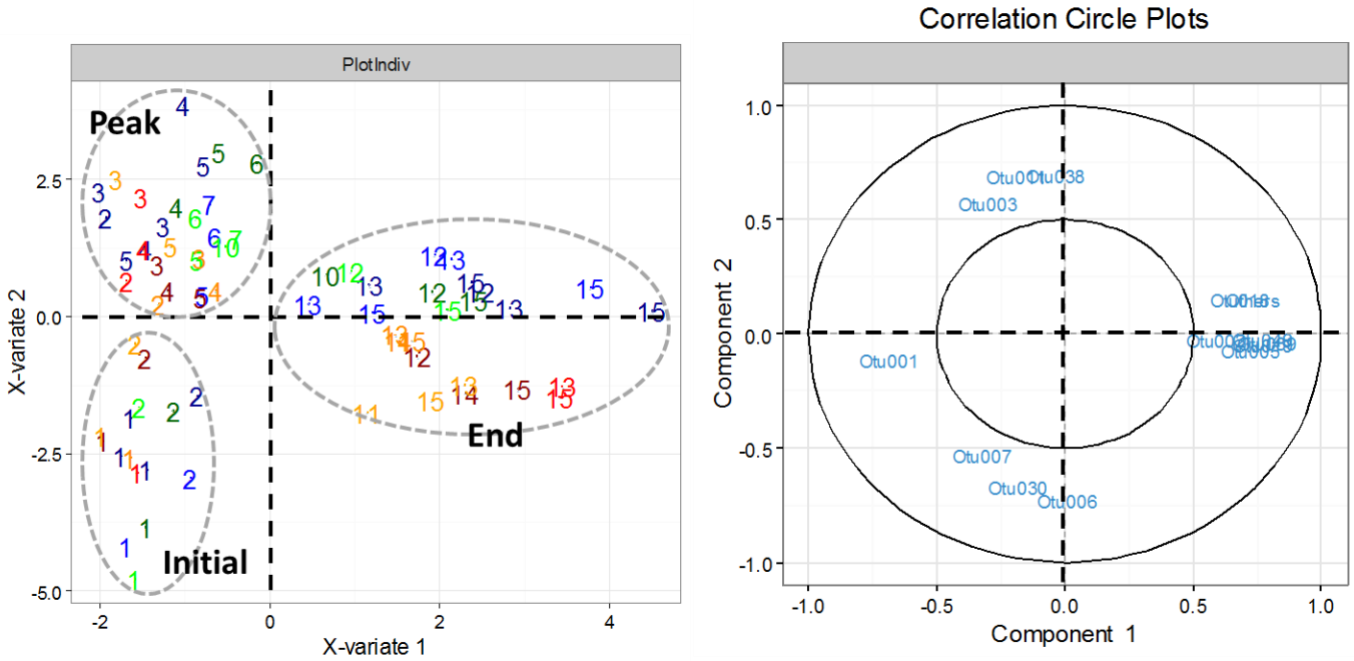


Figure 7. Sparse Partial Least Square Discriminant Analysis (sPLS-DA) of rRNA expression data, performed with degradation phases (initial, degradation peak and end of degradation) as discriminant factor. On the left, plot colors correspond to treatment A (blue), B (red), C (green) and D (orange).

Numbers indicate sampling day of each point. On the right, corresponding correlation circle plot of the selected variables.

Table 2. rRNA expression level (relative to the total rRNA) of sPLS-DA selected OTUs, according to LC degradation phases. Percentages correspond to average level for each phase. Taxonomy at the Class level and at the best level available is given for each OTU.

			Mean rRNA expression			ANOVA	
OTU	Class	Taxonomy	**Initial**	**Peak**	**End**	P-value	
1	*Bacteroidia*	*Bacteroides*	34,0%	28,8%	16,0%	3,55E-06	***
2	*Bacteroidia*	*Rikenellaceae_RC9*	0,3%	0,6%	6,1%	1,05E-07	***
3	*Clostridia*	*ambigous Lachnospiraceae*	1,3%	13,0%	4,9%	8,67E-06	***
4	*Bacteroidia*	*Bacteroides*	2,7%	8,4%	2,7%	4,98E-07	***
5	*γ-proteobacteria*	*Acinetobacter*	3,9%	5,6%	15,1%	7,66E-11	***
6	*Clostridia*	*Clostridium*	11,9%	3,6%	7,3%	6,87E-06	***
7	*Bacilli*	*Enterococcus*	13,1%	2,5%	1,0%	4,37E-09	***
11	*Bacteroidia*	*Prevotella*	0,0%	0,8%	0,2%	6,31E-07	***
13	*Clostridia*	*Butyrivibrio*	0,1%	0,6%	0,4%	1,23E-02	*
18	*Clostridia*	*Eubacterium*	0,3%	0,84%	2,8%	3,31E-06	***
19	*Actinobacteria*	*Corynebacterium*	0,5%	0,43%	3,6%	6,04E-07	***
28	*Clostridia*	*ambigous Lachnospiraceae*	0,0%	0,7%	0,5%	1,26E-02	*
30	*Bacteroidia*	*Bacteroides*	1,9%	0,4%	0,7%	7,00E-07	***
32	*Bacilli*	*uncl_Lactobacilles*	1,5%	0,5%	0,2%	1,10E-06	***
35	*Clostridia*	*Clostridium*	0,7%	0,3%	0,4%	8,90E-03	**
37	*Clostridia*	*ambigous Lachnospiraceae*	0,3%	0,68%	0,5%	9,91E-03	**
38	*Clostridia*	*Oscillibacter*	0,1%	0,47%	0,3%	6,45E-06	***
39	*β-proteobacteria*	*uncl_Comamonadaceae*	0,1%	0,2%	1,3%	4,20E-12	***
49	*Clostridia*	*uncl_Ruminococcaceae*	0,05%	0,08%	0,33%	3,24E-07	***
		Others	3,9%	5,0%	10,8%	8,36E-09	***

II.5. Discussion

II.5.1 Correlation between community dynamics and fermentation parameters

After 15 days of fermentation, the LC degradation level was similar with all the pretreatments except treatment C, which clearly had a negative effect on the LC transformation. However, maximal degradation rates were higher and occurred earlier with treatments B and D, so the main effect of dry-chemical pretreatments was to increase the lignocellulose degradation rate. The dynamic of the alpha diversity of the communities growing in different pretreated substrates was strongly impacted, particularly with alkali-pretreated substrates. With all pretreated substrates, the richness, Simpson and Shannon diversity indexes decreased during the first days of incubation. Such impact was less

pronounced compared to the other substrates only when wheat straw was milled to 100µm size. Thus, the increase on LC degradation rate, which was the main effect of these pretreatments, seems to be associated at the microbial level, to a decrease on diversity. This lower diversity implies a more unbalanced community, with few dominant species, which could resulted from the stronger growth of more active bacteria involved on LC transformation.

In Bacteria, 16S rRNA expression is associated with protein translation; it could thus be correlated with growth and/or protein synthesis. This marker is frequently used as an estimator of bacterial activity which present greater and quicker variations than 16S rDNA (Kerkhof and Kemp, 1999). Moreover, 16S rRNA profiles allow the identification of actually active OTU in the system while 16S rDNA data described the whole community diversity, irrespective of its activity (Dar et al., 2007). Indeed, in this study 16S rDNA profiles appeared almost stable compared to 16S rRNA ones (Supplementary data). In this study, our analysis is based on 16S rRNA profiles which better reflect the community's activity.

II.5.2 Cyclic 16S rRNA profiles

The microbial communities were well separated according to LC degradation phase, as showed by PCA analysis of 16S rRNA sequencing data. Similar results were also observed using more dedicated tools such as weighted-Unifrac distance, which take phylogenetic distance between OTUs into account and Principle Coordinate Analysis (Lozupone et al., 2011). These methods are generally used to regroup communities according to different substrate or environmental conditions, but in this study they were used to follow the temporal dynamic of the microbial community throughout LC transformation. Weighted-Unifrac distances and PCoA ordination displayed similar results for all pretreatments showing that RWS microbial consortium presented a cyclic behavior; it reflects RWS capacity to return to its initial 16S rRNA composition at the end of the fermentation. This cyclic behavior has two implications. Firstly, it means that the bacterial consortium was highly stable; despite its composition change along the LC degradation process, once the LC degradable fraction is consumed, the community goes back to its initial composition which would react again in the same way that the initial inoculum. RWS microbial consortium used in this study was enriched by a sequential batch process until diversity of final points was stabilized. Indeed, such stable diversity was accompanied by a stable functioning, corresponding to a similar LC degradation level (Lazuka et al., 2015). To our knowledge, no previous study has reported

such cyclic diversity dynamics of 16S rRNA transcripts. Secondly, throughout the LC degradation process and irrespectively of the substrate pretreatment applied, the maximal distances between communities to the initial community composition was observed during the first days of LC transformation, corresponding precisely to the most active LC transformation period. It suggests that community's diversity at the end of LC fermentation did not represent the most active community for LC deconstruction. It is important information for studies interested on enriching lignocellulolytic microbial communities which generally based the selection and characterization of the most LC-active microorganisms based on the characterization of community obtained at the end of the LC transformation process ((Feng et al., 2011; Peng et al., 2010). A more efficient microbial community enrichment could probably be obtained by targeting the microorganisms active when the highest degradation rates are observed.

II.5.3 OTUs specific of a pretreatment

Several OTUs with differential 16S rRNA expression in the different pretreated substrates were identified by sPLS-DA statistical analysis. Such OTUs could partially explain the substrate pretreatment effects on community composition. Unfortunately, half of these OTUs were not assigned at the genus level, so it is difficult to identify the role they could play in the community. Nevertheless, *Bacteroides* OTU_1 was dominant in all treatments, but its activity was particularly high for NaOH dry-mechanical pretreatments. *Bacteroides* are known to be a mammalian digestive commensals, and has been frequently described as a xylanolytic bacteria displaying a wide battery of sugar transporters (Tomomi Nishiyama, 2009). The activity of *Clostridium* OTU_6, as those of *Clostridiales* OTU_{15}, appeared to be very specific to the NaOH dry-mechanical pretreatments. *Clostridium* relatives are often cellulose degraders and can use various sugars including hemicellulose components (Thomas et al., 2014). Moreover, numerous *Clostridiales* genomes contain genes coding for xylanases and cellulosomes, which are particularly important in ruminal cellulose digestion. At the opposite, *Enterococcus* OTU_7 presented a lower activity with pretreated straw. This OTU was mainly observed during the first days of LC degradation. *Enterococcus* are widespread digestive commensals, and some of them are able to grow on cellulose as only carbon source (Ramsey et al., 2014). *Enterococcus* are known to tolerate very large panels of pH, temperature and substrate conditions so it is not clear why they were active only in the first days of incubation and almost disappeared in all pretreated straws. Finally, *Butyrivibrio* OTU_{17}, was observed in

all pretreated substrates. This genus is known to utilize xylose to preferentially produce butyrate (Forster et al., 1996), which is one of the carboxylates observed during LC transformation (Lazuka et al. 2016).

II.5.4 OTUs related to specific LC degradation phase

Using sPLS-DA to find discriminant OTUs participating to the LC transformation processes enabled the identification of some OTUs related to a specific LC degradation phase. For all pretreatments, *Bacteroides* OTU_1 displayed 2-fold higher activity during initial and active phases of LC transformation than at final points. As mentioned above, *Bacteroides* has been frequently described as a xylanolytic bacteria. OTU_2 was mainly observed at the latest incubation times. This OTU was classified only at the family level to *Rikenellaceae*, but presented 100% identity with sequences of uncultured and unknown rumen bacteria, identified in different studies (Zened et al., 2012). Little is known about this family, with only two described genus that share an anaerobic metabolism and the ability to tolerate acid media (Graf, 2014). Such ability may be related to its increased activity at the end of the incubation when higher VFA concentrations were observed (Lazuka et al.., 2015). An *Actinobacteria* related OTU (*Corynebacterium* OTU_{19}) was also identified as more active at final points for all pretreatments. *Actinobacteria* are rather rare in rumen, where their function are not well characterized (Suľák et al., 2012). However, in soils they are often involved in organic matter decomposition and thus are able to display various enzymatic activities related to lignocellulose degradation (Negassa et al., 2015).

Surprisingly, OTU_5, which was highly active at final times of incubation, was assigned to *Acinetobacter* which is a genus known as strictly aerobic. However, the sequence of this OTU presented 99% identity with *Acinetobacter* sp. HR7, recently isolated from a Korean cattle rumen (Chang et al., 2015). This genus was also observed several times in rumen (Mao et al., 2013) but it is not clear if the isolate is able to grow without oxygen. *Acinetobacter* has been described as linked to nitrate removal in anaerobic processes, but also with acetate degradation in fluctuating anaerobic/aerobic environments (Su et al., 2015). *Acinetobacter* isolates from termite guts have been reported as cellulose-degraders (Pourramezan et al., 2012); so despite its activity in anaerobic environments is unclear, OTU_5 could be involved in the slow degradation of the refractory cellulosic substrate. *Fibrobacteres* were active at the end of the LC degradation process, however they were not significantly characteristic of a

period, as determined by ANOVA analysis. Nevertheless, this phylum is involved in ruminal digestion (Ransom-Jones et al., 2014) but here was not the most active.

II.6. Conclusion

16S rRNA profiles showed a cyclic behavior during LC degradation, so the active community at LC degradation peak was strongly different compared to initial or final communities. RWS, the cow rumen-derived inoculum used in this study, was thus highly stable. Active diversity profiles at LC degradation peak were similar between pretreatments, with a dominance of *Bacteroidia* and *Clostridia*, but increased degradation speed in response to substrate pretreatments was correlated to increased activity of some OTUs (mainly *Bacteroides* and *Clostridium*), that appeared favored by NaOH treatments.

II.7. Acknowledgments

This research was supported by the French National Institute for Agronomical Research (INRA) and the Region Languedoc-Roussillon Midi-Pyrénées. The authors thank the Genomics and Transcriptomics (GeT) platform for their help with sequencing, and acknowledge Cécile Roland and Maïder Abadie for technical support and their help with experiments.

II.8. References

Barakat, A., Chuetor, S., Monlau, F., Solhy, A., and Rouau, X. (2014). Eco-friendly dry chemo-mechanical pretreatments of lignocellulosic biomass: Impact on energy and yield of the enzymatic hydrolysis. Appl. Energy *113*, 97–105.

Bokulich, N.A., Subramanian, S., Faith, J.J., Gevers, D., Gordon, J.I., Knight, R., Mills, D.A., and Caporaso, J.G. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat. Methods *10*, 57–59.

Cao, K.-A.L., González, I., and Déjean, S. (2009). integrOmics: an R package to unravel relationships between two omics datasets. Bioinformatics *25*, 2855–2856.

Chandel, A.K., and Singh, O.V. (2011). Weedy lignocellulosic feedstock and microbial metabolic engineering: advancing the generation of "Biofuel." Appl. Microbiol. Biotechnol. *89*, 1289–1303.

Chang, D.-H., Rhee, M.-S., Jeong, H., Kim, S., and Kim, B.-C. (2015). Draft Genome Sequence of Acinetobacter sp. HR7, Isolated from Hanwoo, Korean Native Cattle. Genome Announc. *3*.

Dar, S.A., Yao, L., Dongen, U. van, Kuenen, J.G., and Muyzer, G. (2007). Analysis of Diversity and Activity of Sulfate-Reducing Bacterial Communities in Sulfidogenic Bioreactors Using 16S rRNA and dsrB Genes as Molecular Markers. Appl. Environ. Microbiol. *73*, 594–604.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. Bioinformatics *27*, 2194–2200.

Feng, Y., Yu, Y., Wang, X., Qu, Y., Li, D., He, W., and Kim, B.H. (2011). Degradation of raw corn stover powder (RCSP) by an enriched microbial consortium and its community structure. Bioresour. Technol. *102*, 742–747.

Forster, R. j., Teather, R. m., Gong, J., and Deng, S.-J. (1996). 16s rDNA analysis of Bufyrivibrio fibrisolvens: phylogenetic position and relation to butyrate-producing anaerobic bacteria from the rumen of white-tailed deer. Lett. Appl. Microbiol. *23*, 218–222.

Graf, J. (2014). The Family Rikenellaceae. In The Prokaryotes, E. Rosenberg, E.F. DeLong, S. Lory, E. Stackebrandt, and F. Thompson, eds. (Springer Berlin Heidelberg), pp. 857–859.

Hendriks, A.T.W.M., and Zeeman, G. (2009). Pretreatments to enhance the digestibility of lignocellulosic biomass. Bioresour. Technol. *100*, 10–18.

Kerkhof, L., and Kemp, P. (1999). Small ribosomal RNA content in marine Proteobacteria during non-steady-state growth. FEMS Microbiol. Ecol. *30*, 253–260.

Lazuka, A., Auer, L., Bozonnet, S., Morgavi, D.P., O'Donohue, M., and Hernandez-Raquet, G. (2015). Efficient anaerobic transformation of raw wheat straw by a robust cow rumen-derived microbial consortium. Bioresour. Technol. *196*, 241–249.

Lozupone, C., Lladser, M.E., Knights, D., Stombaugh, J., and Knight, R. (2011). UniFrac: an effective distance metric for microbial community comparison. Isme J. *5*, 169–172.

Magoč, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinforma. Oxf. Engl. *27*, 2957–2963.

Mao, S.Y., Zhang, R.Y., Wang, D.S., and Zhu, W.Y. (2013). Impact of subacute ruminal acidosis (SARA) adaptation on rumen microbiota in dairy cattle using pyrosequencing. Anaerobe *24*, 12–19.

McMurdie, P.J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PloS One *8*, e61217.

Mikaelyan, A., Köhler, T., Lampert, N., Rohland, J., Boga, H., Meuser, K., and Brune, A. (2015). Classifying the bacterial gut microbiota of termites and cockroaches: A curated phylogenetic reference database (DictDb). Syst. Appl. Microbiol. *38*, 472–482.

Mosier, N., Wyman, C., Dale, B., Elander, R., Lee, Y.Y., Holtzapple, M., and Ladisch, M. (2005). Features of promising technologies for pretreatment of lignocellulosic biomass. Bioresour. Technol. *96*, 673–686.

Negassa, W.C., Guber, A.K., Kravchenko, A.N., Marsh, T.L., Hildebrandt, B., and Rivers, M.L. (2015). Properties of Soil Pore Space Regulate Pathways of Plant Residue Decomposition and Community Structure of Associated Bacteria. Plos One *10*, e0123999.

Peng, G., Zhu, W., Wang, H., Lue, Y., Wang, X., Zheng, D., and Cui, Z. (2010). Functional Characteristics and Diversity of a Novel Lignocelluloses Degrading Composite Microbial System with High Xylanase Activity. J. Microbiol. Biotechnol. *20*, 254–264.

Pourramezan, Z., Ghezelbash, G.R., Romani, B., Ziaei, S., and Hedayatkhah, A. (2012). Screening and identification of newly isolated cellulose-degrading bacteria from the gut of xylophagous termite Microcerotermes diversus (Silvestri). Microbiology *81*, 736–742.

Ramsey, M., Hartke, A., and Huycke, M. (2014). The Physiology and Metabolism of Enterococci. In Enterococci: From Commensals to Leading Causes of Drug Resistant Infection, M.S. Gilmore, D.B. Clewell, Y. Ike, and N. Shankar, eds. (Boston: Massachusetts Eye and Ear Infirmary), p.

Ransom-Jones, E., Jones, D.L., Edwards, A., and McDonald, J.E. (2014). Distribution and diversity of members of the bacterial phylum Fibrobacteres in environments where cellulose degradation occurs. Syst. Appl. Microbiol. *37*, 502–509.

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., et al. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl. Environ. Microbiol. *75*, 7537–7541.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., et al. (2014). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539–539.

Snajdr, J., Cajthaml, T., Valášková, V., Merhautová, V., Petránková, M., Spetz, P., Leppänen, K., and Baldrian, P. (2011). Transformation of Quercus petraea litter: successive changes in

litter chemistry are reflected in differential enzyme activity and changes in the microbial community composition. FEMS Microbiol. Ecol. *75*, 291–303.

Stamatakis, A. (2014). RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics btu033.

Su, J.F., Zheng, S.C., Huang, T.L., Ma, F., Shao, S.C., Yang, S.F., and Zhang, L.N. (2015). Characterization of the anaerobic denitrification bacterium Acinetobacter sp SZ28 and its application for groundwater treatment. Bioresour. Technol. *192*, 654–659.

Suľák, M., Sikorová, L., Jankuvová, J., Javorský, P., and Pristaš, P. (2012). Variability of Actinobacteria, a minor component of rumen microflora. Folia Microbiol. (Praha) *57*, 351–353.

Thomas, L., Joseph, A., and Gottumukkala, L.D. (2014). Xylanase and cellulase systems of Clostridium sp.: An insight on molecular approaches for strain improvement. Bioresour. Technol. *158*, 343–350.

Tomomi Nishiyama, A.U. (2009). Bacteroides graminisolvens sp. nov., a xylanolytic anaerobe isolated from a methanogenic reactor treating cattle waste. Int. J. Syst. Evol. Microbiol. *59*, 1901–1907.

Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K.-H., Ludwig, W., Glöckner, F.O., and Rosselló-Móra, R. (2008). The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst. Appl. Microbiol. *31*, 241–250.

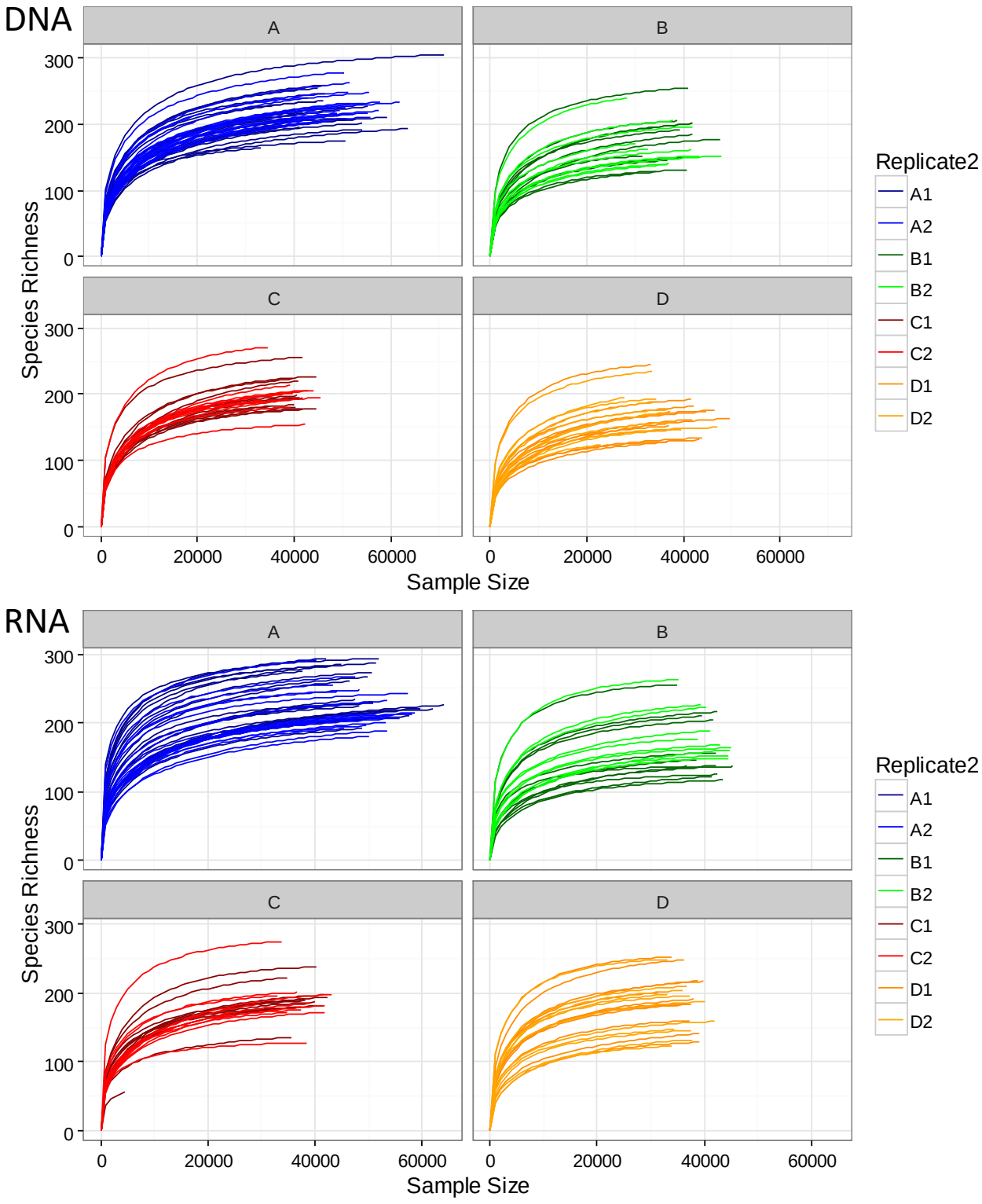
Zened, A., Combes, S., Cauquil, L., Mariette, J., Klopp, C., Bouchez, O., Troegeler-Meynadier, A., and Enjalbert, F. (2012). Microbial ecology of the rumen evaluated by 454 GS FLX pyrosequencing is affected by starch and oil supplementation of diets. FEMS Microbiol. Ecol. n/a–n/a.

Zhao, Y., Wang, Y., Zhu, J. y., Ragauskas, A., and Deng, Y. (2008). Enhanced enzymatic hydrolysis of spruce by alkaline pretreatment at low temperature. Biotechnol. Bioeng. *99*, 1320–1328.

II.9. Supplementary data

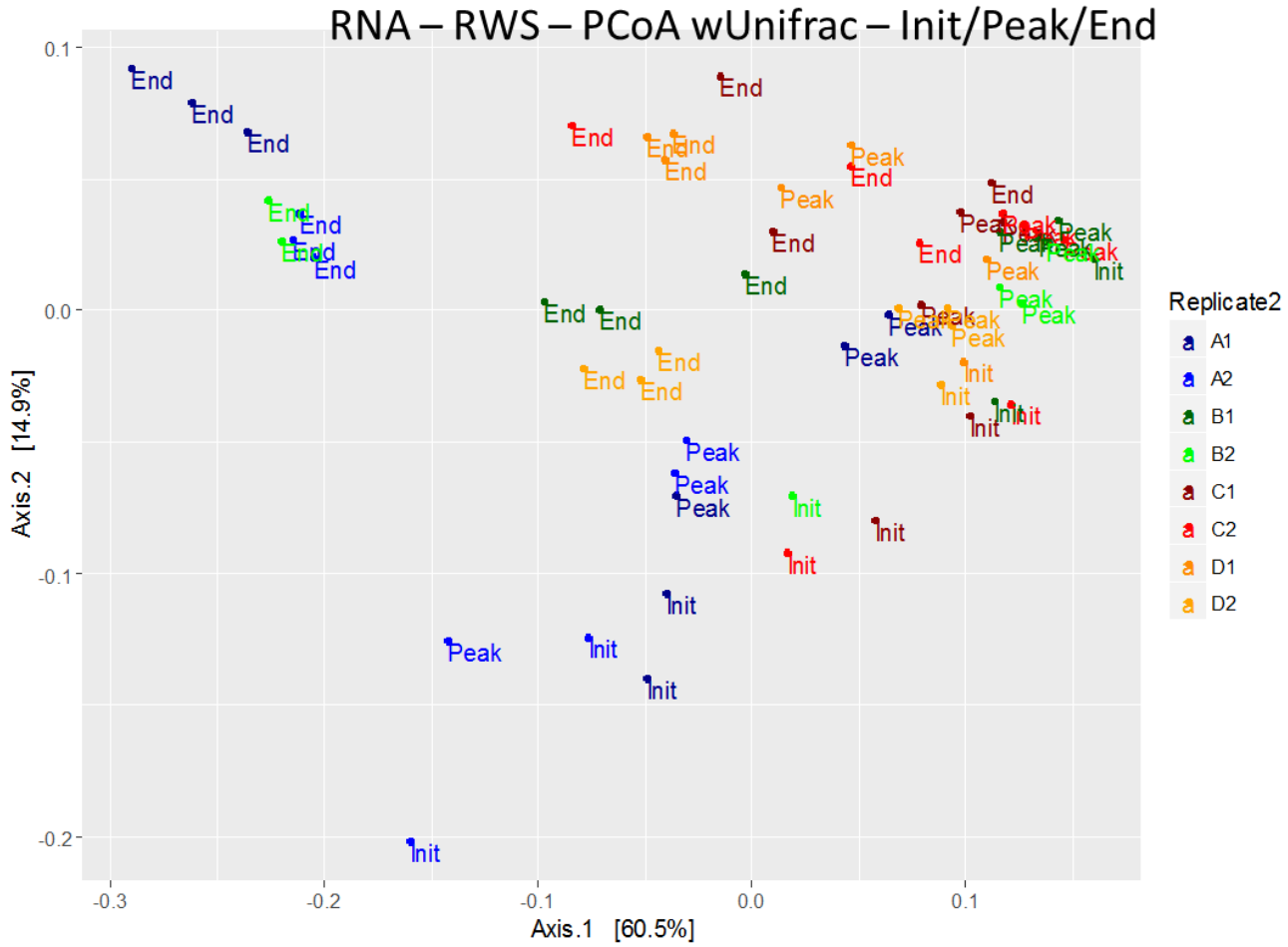
Supplementary Data 1

Rarefaction curves for DNA and RNA sequencing data for each treatment. Each curve corresponds to a sequenced sample. All samples were close to saturation so the sequencing depth was sufficient to describe the communities. The only exception was for the t=0 RNA sample of treatment C1, which had only 4412 sequences.



Supplementary Data 2

Principle Coordinate Analysis plots for all treatments, based on weighted-Unifrac distances and RNA data. A: 2mm, B: 2mmNaOH, C: 100µm, D: 100µmNaOH. Only points corresponding to initial state, degradation peak and final state were kept. A PERMANOVA analysis was performed using type of point (Init, Peak or End : degrad_sep), Treatment and Replicate as factors. Percentage of wUnifrac distances explained is detailed in the R2 column. Pseudo P-values are indicated in the Pr column.



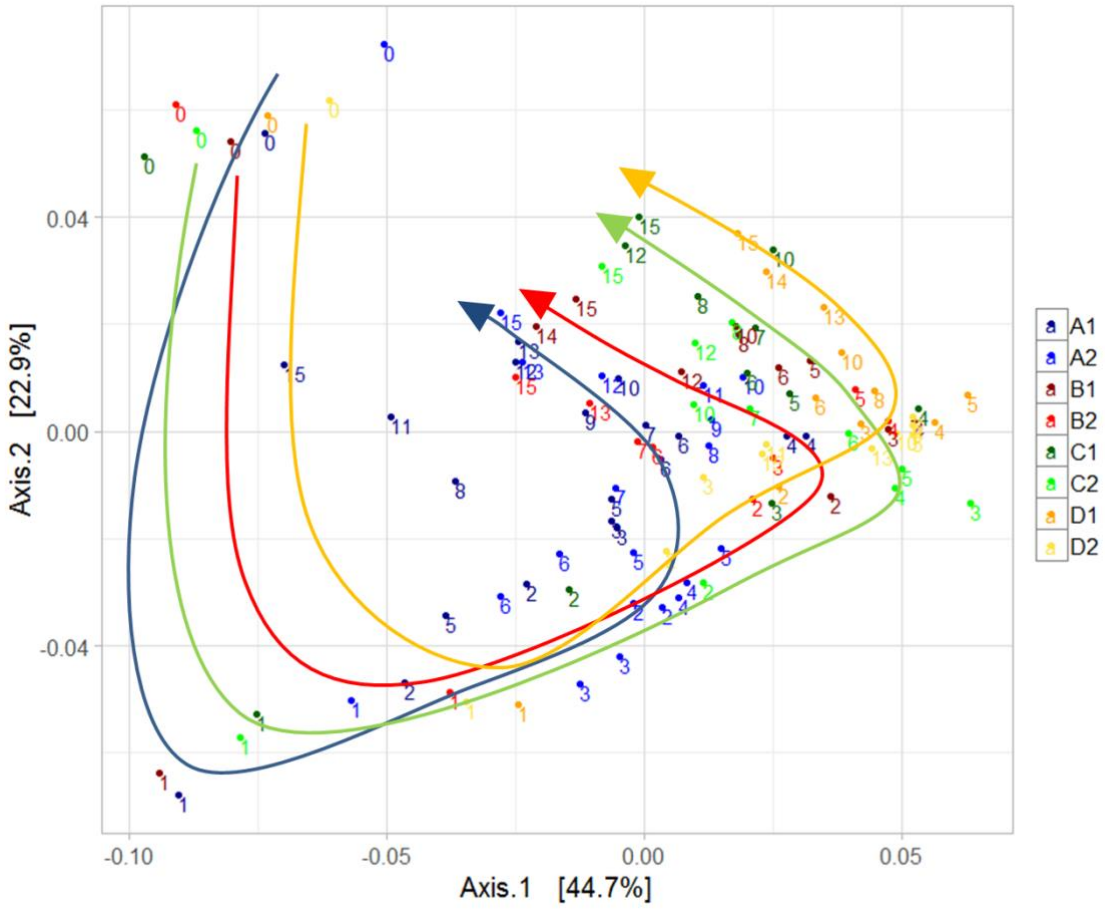
PERMANOVA

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)	
degrad_sep	2	0.51759	0.258797	46.895	0.36135	1,00E-04	***
Treatment	3	0.42424	0.141414	25.625	0.29617	1,00E-04	***
Replicate	4	0.08505	0.021261	3.853	0.05937	3,00E-04	***
degrad_sep:Treatment	6	0.16271	0.027118	4.914	0.11359	1,00E-04	***
Residuals	44	0.24282	0.005519		0.16952		
Total	59	1.43240			1.00000		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Supplementary Data 3

Principle Coordinate Analysis based on 16S RNA weighted-Unifrac distances. Numbers correspond to point sampling days, colors to treatments and replicates A (2mm), B (2mm-NaOH), C (100µm) and D (100µm-NaOH). Arrows underline the cyclic behavior for each treatment.



III. Conclusion du chapitre

Le prétraitement du substrat par imprégnation à la soude a un impact positif sur sa vitesse de dégradation de la lignocellulose par le consortium RWS. Cette augmentation de vitesse est par ailleurs corrélée à des variations du profil d'activité des communautés bactériennes, et il est possible d'identifier des taxons qui répondent différemment selon les prétraitements. On observe globalement une forte diminution de la diversité, due à une augmentation de la proportion des *Bacteroidia* (plus particulièrement du genre *Bacteroides*) et *Clostridia*, qui deviennent très majoritaires.

Le profil d'activité des communautés bactériennes présente par ailleurs un comportement cyclique au cours du temps quel que soit le traitement utilisé. Le retour vers l'état initial en fin de dégradation traduit la forte stabilité de la communauté, indépendante du prétraitement utilisé. Cependant, cette dynamique cyclique implique également que lors du pic de dégradation, les populations métaboliquement les plus actives sont différentes de celles actives à l'état initial ou final. En se basant sur les données de dégradation pour distinguer les points appartenant à trois phases, initiale, final ou pic de dégradation, il est possible d'identifier des taxons dont l'activité métabolique est différente en fonction de la phase de dégradation. Avec un substrat prétraité, ceux-ci sont principalement affiliés *Bacteroides* et *Clostridia*.

La dispersion des communautés microbiennes en fonction de leur distance Unifrac pondérée est majoritairement due à la phase de dégradation dans laquelle elles sont observées, les communautés étant ensuite séparées en fonction du prétraitement. Les communautés microbiennes sont donc plus impactées par l'évolution du substrat au cours de la dégradation que par le prétraitement, même si celui-ci a un effet fort au sein de chaque phase considérée.

CHAPITRE VI

EXPLORATION DU POTENTIEL DE DEGRADATION DE FLORES INTESTINALES DE TERMITES

CHAPITRE VI : EXPLORATION DU POTENTIEL DE DEGRADATION DE FLORES INTESTINALES DE TERMITES

I. Introduction

Dans les chapitres précédents, un consortia bactérien (RWS) présentant de très bonnes capacités lignocellulolytiques a été sélectionné à partir de rumen bovin. Sa stabilisation a abouti à une communauté avec une diversité réduite et composée principalement de *Bacteroidetes* et *Firmicutes*. Sa composition au niveau du phylum est assez similaire à l'inoculum de départ. Les performances de dégradation et production obtenues ont ainsi montré que l'utilisation d'un inoculum provenant d'un écosystème naturel capable de digérer la lignocellulose pouvait être une stratégie payante. Cependant, les taux de dégradation atteints sont encore en dessous de ce que les écosystèmes naturels sont capables d'atteindre, même en utilisant des substrats prétraités. Le rumen bovin, précédemment testé et acclimaté, n'est de plus pas l'écosystème présentant le meilleur taux de dégradation.

Les insectes, et notamment les termites, atteignent des taux de dégradation supérieurs, jusqu'à 74-99% pour la cellulose et 65-87% pour l'hémicellulose (Brune, 2014). Ils sont également capables de digérer des substrats très lignifiés, et jouent ainsi un rôle central dans le cycle du carbone des régions tropicales et sous-tropicales. De plus, ils présentent des microbiotes intestinaux très différents des écosystèmes digestifs de mammifères ou des écosystèmes du sol. Les termites supérieurs ont ainsi un microbiote principalement bactérien, et dominé par les *Spirochaetes*, *Fibrobacteres* et d'un troisième phylum peu caractérisé. Cette diversité nouvelle a déjà été étudiée pour la recherche de nouvelles enzymes lignocellulolytiques, mais peut également être une source d'inoculum pour une approche de culture en fermenteurs.

L'inoculation de fermenteurs par le microbiote intestinal de différentes espèces de termites peut permettre d'une part, de vérifier si comme pour le rumen, il est possible de maintenir une activité de dégradation tout en stabilisant une communauté bactérienne, et d'autre part, de suivre l'évolution de la diversité bactérienne pendant son acclimatation aux conditions de culture. D'autre part, cette seconde source d'inoculation peut nous permettre d'évaluer l'impact de l'origine et de la composition de l'inoculum sur les performances de dégradation de la biomasse, qui est une question fondamentale en écologie microbienne.

Ce chapitre décrit donc l'exploration des capacités de dégradation de lignocellulose en fermenteur de quatre microbiotes intestinaux de termite supérieurs, mis en culture dans des fermenteurs anaérobies contenant de la paille de blé comme seule source de carbone, et le suivi de la diversité microbienne associée, obtenu par séquençage de l'ADNr 16S.

II. Exploration du potentiel de microbiote intestinal de termite comme biocatalyseur de la bioconversion de lignocellulose

Ce chapitre sera prochainement soumis à « Environmental Microbiology » sous le titre :

Exploring the potential of termite gut microbiota as biocatalyst for lignocellulose bioconversion

Lucas Auer[1,2,3], Adèle Lazuka[1,2,3], David Sillam-Dusses[4,5], Edouard Miambi[5], Claire Dumas[1,2,3], Michael O'Donohue[1,2,3], Guillermina Hernandez-Raquet[1,2,3*]

[1)] Université de Toulouse, INSA, UPS, LISBP, 135 Avenue de Rangueil, F-31077 Toulouse, France
[2)] INRA, UMR792 Ingénierie des Systèmes Biologiques et des Procédés, F-31400 Toulouse, France
[3)] CNRS, UMR5504, F-31400 Toulouse, France
[4)] University Paris 13 - SPC, LECE, EA4443, 99 avenue Jean-Baptiste Clément, F-93430 Villetaneuse, France
[5)] IRD – Sorbonne University, IEES, U242, 32 avenue Henri Varagnat, F-93140 Bondy, France

II.1. Abstract

Termites display very high lignocellulose degradation capacities being able to degrade wood with high lignin content. Their digestion process mainly relies on a symbiotic relationship with microorganisms. The gut microbiome of higher termites is only colonized by bacteria, whose taxonomy differs significantly from well-known lignocellulolytic communities of soil or mammal's digestive systems. Termite gut microbiomes have been studied for enzyme discovery purpose; they also represent a potential source of microorganisms to catalyze bioconversion processes such as the carboxylates platform. In this work, the gut microbiomes of four termite species were studied for their lignocellulolytic and carboxylates production potential in controlled bioreactors. All of the termite gut microbiomes studied here presented high lignocellulose conversion rates degrading up to 45% wheat straw. *Nasutitermes ephratae* gut-microbiome was the best of the tested ecosystems, displaying a high enzymatic activity, degradation rate and carboxylates production. 16S rRNA gene sequencing revealed important changes on the community composition between inocula and the final reactor's communities. *Spirochaetes* and *Fibrobacteres*, dominant in termite gut communities, were replaced by *Firmicutes* and *Proteobacteri*a, indicating that low-abundance termite-gut species were selected in the lignocellulolytic bioreactors. This work demonstrates the potential of termite-gut microbiomes to degrade wheat straw in controlled conditions.

II.2. Introduction

With more than 200 billion tons of non-food lignocellulosic biomass produced yearly, lignocellulose represents the most abundant and promising source of renewable carbon on Earth (Chandel and Singh, 2011). The bioconversion of lignocellulosic residues into biofuels or chemicals of industrial interest is thus receiving much attention to reduce our fossil carbon dependence. Lignocellulose is an association of microfibrils of cellulose, hemicelluloses and lignin polymers and a mixture of other secondary components (Lynd et al., 2002; Barakat et al., 2013). Each polymer presents its own complex physicochemical structure and its degradation involves a large panel of cellulases, hemicellulases and lignin-degrading enzymes. Deconstructing such a complex substrate is therefore an enzymatic challenge. Indeed, the major obstacle to the development of lignocellulose bioconversion lies in the recalcitrant structure of lignocellulosic substrates which makes the depolymerization of plant biomass inefficient and non-profitable at an industrial scale.

Nevertheless, lignocellulose degradation is at the basis of carbon recycling in almost all terrestrial and aquatic ecosystems. In Nature, lignocellulose digestion is performed by various organisms, from bacteria to animals, which are valuable resources of deconstruction strategies useful to improve biomass bioconversion processes. In soils, the most studied system for plant biomass degradation, lignocellulose is mainly degraded by the action of aerobic fungi and bacteria able to produce extracellular hydrolytic enzymes to attack the polymeric lignocellulosic matrix. In such ecosystems, lignocellulose is partially mineralized, producing CO_2 and H_2O, accompanied by the formation of recalcitrant humic substances (Deacon, 2005). Due to the high hydrolytic capacity of fungal enzymes, single strains of fungi are currently the main sources of the hydrolytic enzymes used in the sugar platform for bio-ethanol production. In animals, although herbivores and omnivores depend on plant biomass as source of food, most of them lack of endogenous lignocellulolytic enzymes: their capacity to break down and utilize lignocellulose relies on symbiotic relationships with microorganisms (Smant et al., 1998). Today, the symbiotic lignocellulosic system which is better described is the rumen ecosystem of ruminant animals where bacteria, archaea, protozoa and anaerobic fungi are involved in lignocellulose deconstruction (Hobson, 1998). The main products of rumen fermentation by microorganisms are short chain carboxylates (volatile fatty acids, VFA) and methane which represent interesting biotechnological products. The biotechnological application of ruminant fermentation has thus already been

assessed e.g. as a potential inoculum to enhance methane production from diver lignocellulosic feedstocks (Yue et al., 2007).

Without doubt, the best lignocellulose degraders are insects, and particularly termites, which utilize crystalline cellulose and very lignified and recalcitrant substrates, including wood, as their main food source (Breznak and Brune, 1994). Termites are able to remove 74-99% cellulose and 65-87% hemicellulose of wood, being more efficient than ruminants (Brune, 2014). Termites represent thus a promising source for plant cell wall degrading microorganisms and enzymes. Termites belong to the order Isoptera; they are classified in lower (families Masto-, Kalo-, Hodo-, Rhino- and Serritermitidae) and higher (family Termitidae) termites displaying different mechanisms for lignocellulose digestion. Lower termites degrade lignocellulose thanks to a complex symbiotic interaction between the host, eukaryotic flagellates and bacteria (Ni and Tokuda, 2013). Higher termites belonging to *Macrotermitinae* subfamily subsist on lignocellulose diet thought to an external symbiosis with *Basidiomycetes*. Such symbiotic fungi break down lignocellulose in fungal gardens prior to the ingestion by termites (Brune, 2014). In other higher termite species, lignocellulose digestion is realized by the activity of mutualistic hindgut microbiota exclusively constituted by prokaryotes displaying enzymatic activities associated mainly to the wood fibers (Slaytor, 1992; Tokuda and Watanabe, 2007). For several species of both lower and higher termites, the lignocellulolytic microbial activity seems to be coupled to the action of endogenous cellulases produced by the host (Lo et al., 2010; Watanabe et al., 1998). In contrast to the rumen and cellulolytic soil bacterial communities which are dominated by *Firmicutes* and *Bacteroidetes*, the particle-associated bacteria in termite gut are mainly constituted by *Spirochaetes* and *Fibrobacteres* (Mikaelyan et al., 2014). Such contrasting characteristics of termite microbiota may certainly reflect distinct degradation mechanisms and conversion capacities.

Termites have been the subject of numerous researches due to their dietary habits mainly based on lignocellulose digestion. Such studies have been focused on the characterization of their *in-vivo* and *in -vitro* hydrolytic enzyme activities (König et al., 2013). Recently, various function-based or sequence-based metagenomic studies have intended to identify lignocellulose genes from termite gut microbiomes (reviewed recently by (Scharf, 2015). However, only little information exists on the potential of exploiting termite gut microbiota as a source of bacterial inoculum useful for lignocellulose bioconversion for biotechnological applications.

As previously mentioned, the main products of biomass deconstruction by microbiota of termite guts are carboxylates which are assimilated by the host. Thus, termite gut could be an interesting source of efficient biomass degrading bacterial communities useful in the carboxylate platform. This platform, beside the sugar-to-ethanol platform, is one of the most promising approaches for biomass valorization (Agler et al., 2011; Holtzapple et al., 1999). In the carboxylate platform, lignocellulosic biomass is deconstructed by anaerobic mixed bacterial communities, under non-sterile conditions, to produce carboxylates, mainly acetic, propionic and butyric acid, which are interesting products for further biological or chemical transformations (Agler et al., 2011; Kleerebezem and van Loosdrecht, 2007).

Numerous attempts have been realized to obtain efficient lignocellulolytic communities from soils (DeAngelis et al., 2012; Feng et al., 2011), compost (Guo et al., 2010; Reddy et al., 2013), marine sediments (Hollister et al., 2011) or extreme environments (Cope et al., 2014). However, lignocellulose conversion yields are still not sufficient to be profitable, and little is known about the microbiology behind the process. To our knowledge, no research has been realized to exploit termite gut microbiome in controlled bioreactors for carboxylate production. Therefore, the present study is a comparative analysis of microbiomes from four foregut termite species to produce carboxylates when implemented in controlled bioreactor using wheat straw as sole carbon source. This assessment was coupled to the analysis of substrate, products and deployed enzymatic activities. Finally, we characterized the microbial diversity by 16S rRNA gene sequencing, to compare the diversity of initial gut samples and investigate their changes after culturing in lignocellulose batch reactors.

II.3. Results

II.3.1 Reactor performances: degradation and products

Lignocellulose degradation capacities of the gut microbiota from four different species of higher termites was assessed in duplicate reactors per termite species (named r1 and r2) containing wheat straw as sole carbon source. Reactor inocula were standardized to the same number of guts per termite species (500 guts). It corresponded to similar DNA quantities (about 150 µg) and similar number of 16S rRNA gene copies (~$1.5x10^{12}$) per inoculum for *Nasutitermes ephratae, N. lujae*, and *Microcerotermes parvus*. However, lower values were measured for *Termes hospes* (67.9±0.7µg and 1.7±0.7 $x10^{11}$ 16S rRNA copies,Table 1).

Table 1: Microbial biomass concentration estimated with 16S rRNA gene copies measured at the beginning and end of incubations. Values are detailed for each inoculum and biological replicates.

	16S rRNA gene copies/µL			
Species	initial		final	
	r1	r2	r1	r2
M. parvus	5,01E+06	4,34E+06	5,40E+06	1,39E+07
N. ephratae	3,87E+06	3,70E+06	8,61E+06	2,34E+07
N. lujae	2,79E+06	3,88E+06	1,65E+07	1,73E+07
T. hospes	4,31E+05	1,40E+06	1,08E+07	7,83E+06

After 20 days of incubation on anaerobic batch reactors, lignocellulose degradation varied from 26 to 49% for the different inocula and replicates. The highest wheat straw degradation of 45.2±5 % was obtained with *N. ephratae* inoculum, followed by *N. lujae*, with a mean degradation of 37.1± 4.3%, while *M. parvus* and *T. hospes* displayed 30±5% and 31±3.7% degradation, respectively (Figure 1A). Irrespectively of the inoculum, the main products of lignocellulose transformation were carboxylates (VFA, Figure 1B) and CO_2 (data not shown). Minor methane production was detected in some of the reactors but it was immediately inhibited by BES addition. During the first days of culture, a minor H_2 production was also detected but stopped by itself (data not shown). VFA accumulation varied from 2.2 to 5.8g.L^{-1} for the different termite gut inocula; the highest VFA production was obtained with gut microbiota from *N. ephratae* (Figure 1B). For all the termite species tested, VFA was mainly constituted by acetate (>85%) and small amounts of propionate and butyrate, except one of the *M. parvus* replicates which produced about 20% of propionate. It should be noted that VFA production by *N. lujae* and *M. parvus* displayed high variability between replicated reactors (Figure 1B). Despite VFA production kinetics presented some variations, in all reactors a plateau was reached in 20 days showing that the lignocellulose degradation process was over. The observed VFA concentrations in the different reactors were consistent with the corresponding values of lignocellulose degradation and were consistent with theoretical lignocellulose conversion yields (Lazuka et al., 2015). The non-inoculated control reactor showed neither VFA nor gas production throughout the incubation period; in consequence, the residual lignocellulose content remained unchanged (data not shown).

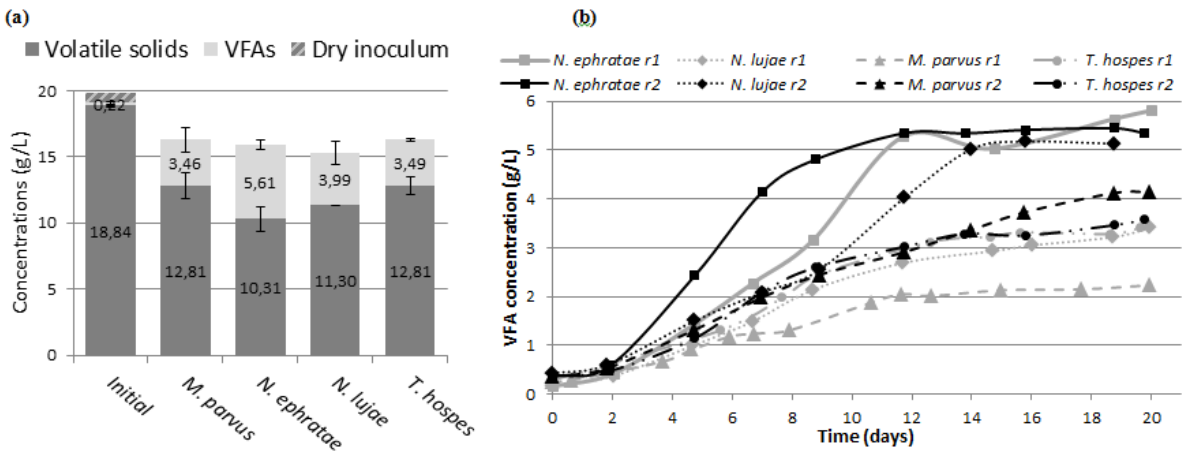


Figure 1: VFA kinetics (a) and degradation and products (b) during fermentation. Sterile wheat straw was inoculated with 500 termite guts and cultured for 20 days. Errors are standard deviations of the two biological replicates.

Enzymatic activities related to lignocellulose degradation were measured on the r2 duplicates at the end of the incubation (Figure 2). Free- and cell-bound activities were detected for both xylanase and cellulase activities whereas exoglucanase and β-glucosidase activities were not detected. While *N. ephrate*, *M. parvus* and *N. lujae* displayed xylanase activities higher than 1500UA.mL^{-1}, *T. hospes* showed a low xylanase activity (600 UA.mL^{-1}). For CMCase activity, a similar profile was observed but lower values of activity were measured in all reactors: 62-71 UA.mL^{-1} CMCase activity for *N. ephrate, M. parvus and N. lujae* while only 40 UA.mL^{-1} were measured for *T. hospes*. For all termite species tested, more than 60% of the xylananse and cellulose activities were cell-bounded.

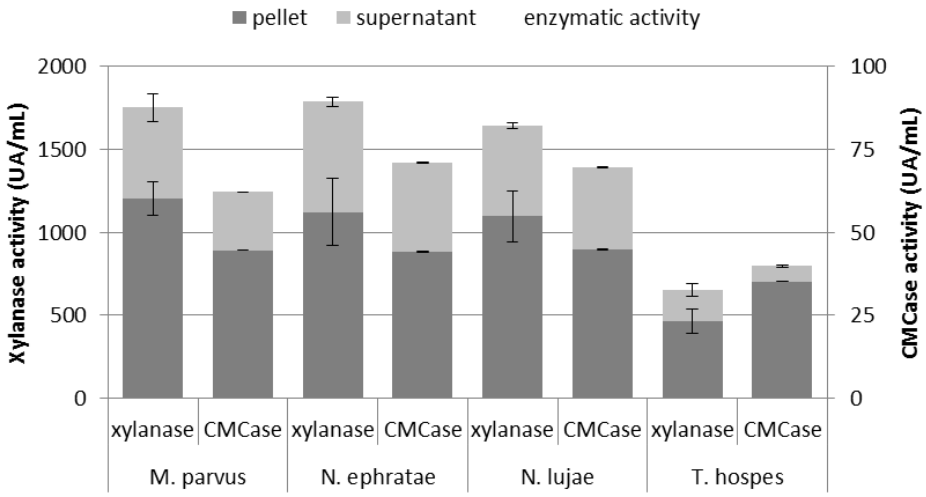


Figure 2: Xylanase and CMCase activities at final points for the four inocula. Supernatants reflect extracellular enzymatic activities whereas pellets reflect cell-bound activities. Errors bars are standard deviations between technical replicates.

II.3.2 Diversity analysis changes of termites gut inocula after lignocellulose degradation

To assess the microbial diversity of termite guts and termite-gut inoculated reactors, V3-V4 16S rRNA gene region was sequenced from genomic DNA samples from the initial termite gut and the end-point of the lignocellulose reactors. A total of 870,449 pair-end reads were successfully assembled in sequences of about 450bp length, with an average of 54,400 reads per sample. After filtering and chimera removal, more than 20,000 high quality sequences per sample remained for further analysis. Sequence clustering yielded a total of 8,794 bacterial operational taxonomic units (OTUs) at 97% sequence similarity threshold. OTUs presenting less than 0.005% of total sequences were removed, considering this value as a detection threshold, resulting on 671 final OTUs. Rarefaction curves, based on normalization by subsampling at 15,000 sequences per sample, showed that with exception of *N. lujae* r1, all samples were close to saturation (Figure S1, Supplementary data). Communities were thus sufficiently sampled to enable us to estimate the actual community diversity and richness.

II.3.3 Gut flora composition

The average number of OTUs observed was 150, 181, 192 and 292 for gut bacterial communities from respectively *T. hospes*, *N. ephrate*, *N. lujae and M. parvus* (Table 2). According with Shannon's and Simpson's reciprocal index *T. hospes* displayed the highest diversity while the other termite gut microbiotes displayed similar levels of diversity.

Table 2: alpha diversity indexes of gut inocula and final reactor communities

	M. parvus				*N. ephratae*				*N. lujae*				*T. hospes*			
	gut1	gut2	r1	r2	gut1	gut2	r1	r2	gut1	gut2	r1	r2	gut1	gut2	r1	r2
ObsOTUs	154	147	63	52	194	169	67	58	196	186	116	44	304	281	61	63
Shannon	3,40	3,31	2,48	1,79	3,43	3,46	2,49	2,10	3,36	3,41	2,20	2,01	4,96	4,94	2,22	2,54
Simpson	12,3	10,1	7,7	3,9	13,2	12,9	7,0	5,4	13,0	12,9	4,9	5,8	64,4	65,7	5,6	7,63

Dominant OTUs were defined as those displaying at least 2% abundance in at least one sample. From the 34 dominant OTUs detected in all the gut samples, only one was shared between three species (*N. lujae*, *N. ephratae* and *M. parvus*), one was shared between *M. parvus* and *T. hospes*, while 9 OTUs were shared between guts of the two *Nasutitermes* species (Supplementary data Table S1). Thus, the majority of dominant OTUs were specific of a given host, and the shared OTUs were concentrated between the two close host species of *Nasutitermes*. This observation was also true for OTUs displaying low abundances (<2%; defined as minor OTUs). Indeed, only 8 minor OTUs were shared between the four termite species. Regarding the gut community's make-up, *M. parvus* and the two species of *Nasutitermes* were dominated by *Spirochaetes* (> 55%), followed by *Fibrobacteres* (except for *N. lujae*) and candidate phylum Termite group 3 (TG3; **Figure 3**, inner circles). A totally different profile was observed for *T. hospes* which was dominated by *Firmicutes* (46%), *Spirochaetes* (21%), *Bacteroidetes* (13%) and *Proteobacteria* (12%). Weighted Unifrac distances of OTUs (**Figure 4**) showed that the two *Nasutitermes* flora were closely related while *M. parvus* was closer to the *Nasutitermes* than *T. hospes*. Despite that *Termes* is phylogenetically closer to *Nasutitermes* than *Microcerotermes* (S2 Supplementary data), *Termes* gut microflora was clearly different; it was the only termite species presenting OTUs belonging to *Firmicutes* and *Proteobacteria*.

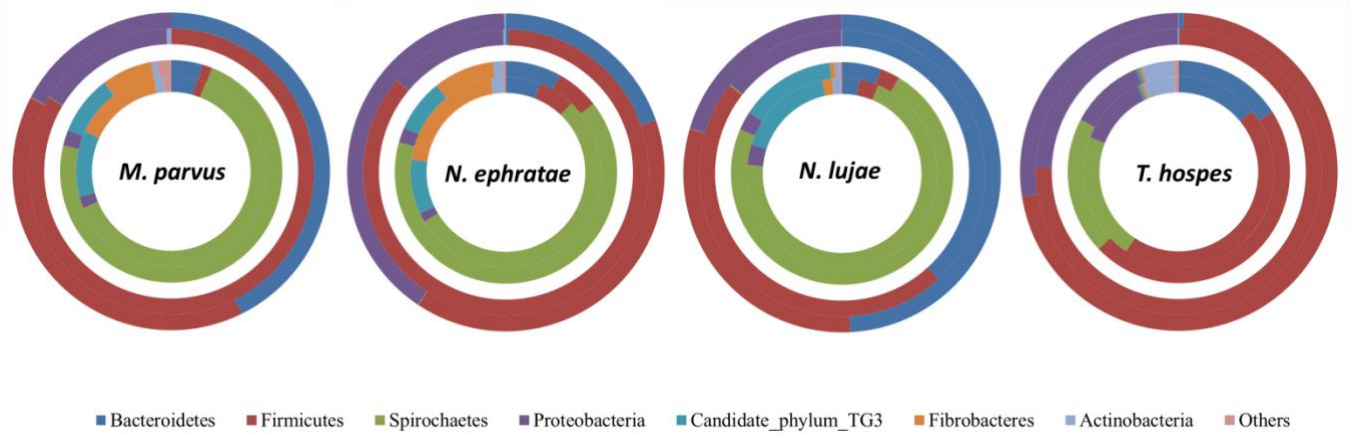


Fig.3. Diversity of the termite-derived microbiota. Phylum level classification of the 16S rRNA genes in guts of selected termites (inner circles) and lignocellulose reactors inoculated with guts of selected termites at the end of the incubation (outer circles). Data for duplicates is in presented. The category 'Others' contains the low abundance phyla *Acidobacteria*, *Chloroflexi*, *Cyanobacteria*, *Chlorobi* and *Deferribacteres*.

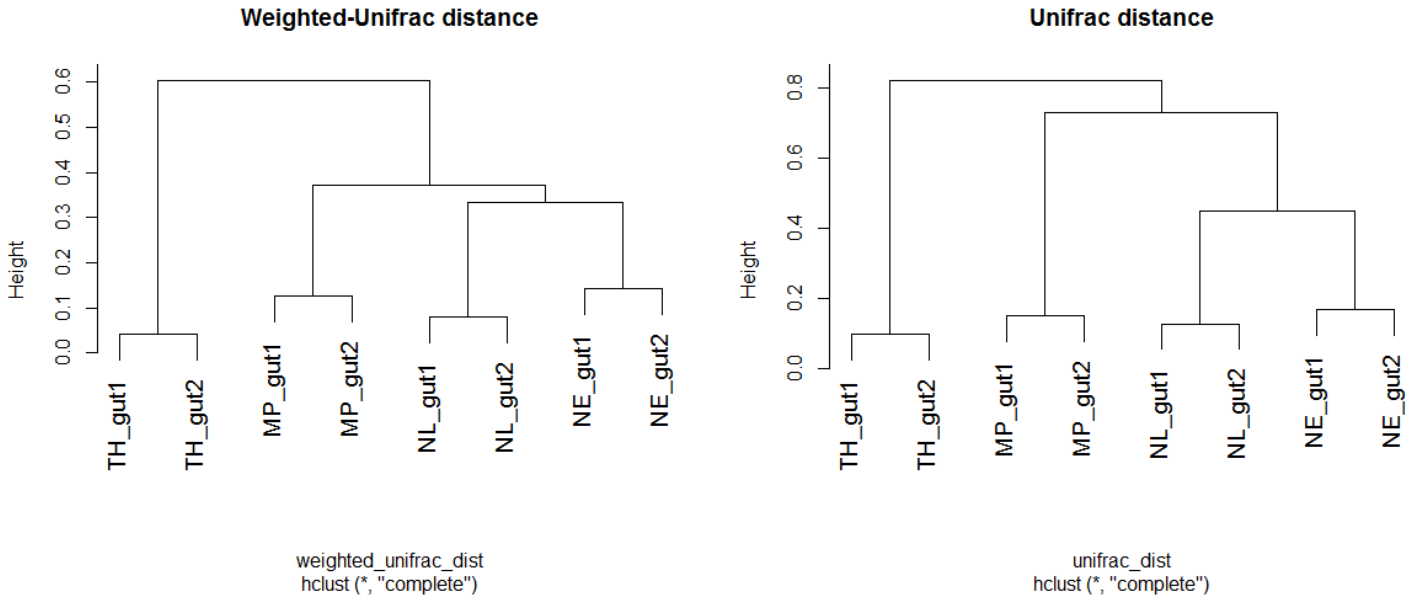


Fig.4. Weighted-Unifrac distances between the four termites' initial communities.

II.3.4 Diversity changes after incubation in lignocellulose reactors

The observed richness (Table 2) using a 97% similarity threshold revealed a significant difference between the number of OTUs observed in the initial termite-gut inocula (between 147 and 304 OTUs) and that found at the end of the incubation in the lignocellulose reactors (less than 67 OTUs). Indeed, half of the observed richness was lost during incubation in bioreactors. Shannon's and Simpson's reciprocal diversity index were higher than 3 and 10, respectively, for all termite-gut inocula, whereas their values were lower than 2.6 and 7.7, respectively, for reactor samples. The observed richness and Shannon's and Simpson's reciprocal index indicated that compared to the original termite-gut inocula, diversity decreased systematically after incubation in lignocellulose reactors.

A majority of sequences (64%) obtained at the end of the incubation in lignocellulose reactors belonged to 59 OTUs which were shared between all libraries. In contrast, 273 OTUs containing 10% of sequences were restricted to a single library. The phylogenetic make-up of reactor end-points was dominated by *Firmicutes* (particularly of the *Clostridia* class), tending to be deficient in *Proteobacteria* and *Bacteroidetes* related OTUs (Figure 3, external circles). Contrary to gut communities, the dominant OTUs in reactor end-points were present in all four libraries irrespective the inoculum source (Table 3). Weighted-Unifrac distances confirmed that the final communities strongly differed from the initial gut communities (**Figure 5**). Although replicates of gut communities from the same termite species clustered

together, replicates of reactor communities were more distant, excepted *T. hospes*. Reactor communities were separated into two subgroups, according to differences at the phylum level. The first group contained communities with a high dominance of *Firmicutes* (over 70%) and almost no *Bacteroidetes*, while the second group was constituted mainly by *Bacteroidetes* (20 to 50%) and *Firmicutes (*30 to 47%).

Table 1 : Relative abundance (%) of the main phyla present in gut and final reactor communities (colored lines). For each phylum, OTU composition is detailed as relative abundance (%) in the given phylum.

	M. parvus				*N. ephratae*				*N. lujae*				*T. hospes*			
	gut1	gut2	r1	r2	gut1	gut2	r1	r2	gut1	gut2	r1	r2	gut1	gut2	r1	r2
Bacteroidetes	4,4	4,6	0,04	42,7	5,9	8,2	0,3	19,7	3,0	5,7	38,2	49,2	14,1	16,0	0,1	0,5
Dysgonomonas Otu002	-	-	-	98,4	-	-	22,2	-	0,7	2,6	99,3	53,4	2,6	-	-	-
Dysgonomonas Otu015	-	-	-	0,02	-	-	-	-	-	-	-	46,1	-	-	-	-
Bacteroides Otu018	-	0,1	-	-	-	-	-	93,8	-	-	-	-	-	-	-	-
Firmicutes	1,5	1,4	83,6	40,0	6,2	6,3	85,8	39,6	3,2	2,9	47,4	30,4	45,3	46,8	75,4	72,0
Clostridium termitidis Otu001	-	-	27,2	60,9	-	-	33,1	34,8	0,2	-	5,8	50,6	0,1	-	17,6	10,7
uncl Lachnospiraceae Otu003	0,4	0,9	8,2	-	0,3	0,7	14,9	2,6	0,6	2,3	44,3	-	0,7	-	45,2	32,9
uncl Lachnospiraceae Otu004	0,9	-	9,2	30,4	-	0,6	21,1	34,5	0,2	-	9,7	37,5	0,1	-	11,2	27,7
Acetanaerobacterium Otu014	-	-	26,8	-	-	-	1,0	-	0,8	-	-	-	-	-	0,3	0,02
Ruminococcaceae Gut_cluster Otu019	-	-	0,1	-	-	-	0,2	-	0,4	-	14,0	-	0,2	-	12,8	0,1
Ruminococcaceae Gut_cluster_7 Otu023	-	-	3,4	-	0,1	-	8,5	-	-	-	3,5	-	0,03	-	-	-
Ruminococcaceae Gut_cluster Otu024	-	-	1,6	1,1	-	-	2,2	7,3	-	-	5,9	0,5	0,04	-	0,4	2,2
Sedimentibacter Otu027	-	-	1,4	0,1	-	-	4,0	2,8	-	-	0,0	-	-	-	3,6	3,0
uncl Clostridia Otu034	-	-	9,1	-	-	-	-	-	-	-	-	-	-	-	-	-
Lachnospiraceae Incertae_Sedis Otu037	-	-	-	-	-	-	-	10,8	-	-	0,03	-	-	-	1,8	2,1
uncl Clostridia Otu042	-	-	3,0	-	-	-	4,2	-	-	-	-	-	-	-	-	-
uncl Lachnospiraceae Otu043	-	-	4,0	0,7	-	-	1,4	0,7	0,2	-	0,2	0,8	-	-	0,03	0,7
Ruminococcaceae Gut_cluster Otu053	-	-	-	-	-	-	-	-	-	-	3,7	-	0,04	-	0,4	3,0
Proteobacteria	1,9	2,3	15,8	17,3	1,6	2,2	13,5	40,5	3,5	2,7	14,2	20,4	11,9	10,8	24,4	27,5
uncl Enterobacteriaceae Otu005	-	-	0,1	38,2	0,4	1,2	21,7	81,2	0,2	0,5	43,1	55,6	0,4	0,7	8,0	14,2
Pseudomonas_2 Otu010	-	-	0,6	32,1	0,4	-	4,8	2,5	0,4	-	21,5	16,5	0,2	-	63,5	33,3
uncl Enterobacteriaceae Otu012	-	-	0,0	12,9	-	-	3,7	1,0	-	-	26,2	22,1	0,2	-	1,4	47,3
Stenotrophomonas Otu028	-	-	44,7	0,3	-	-	24,5	-	-	-	1,0	-	-	-	-	-
uncl Rhodocyclales Otu029	-	-	10,1	-	0,4	-	25,0	-	0,2	-	6,4	-	0,2	-	15,9	-
Escherichia-Shigella Otu031	0,7	5,6	37,1	-	0,9	1,2	17,6	0,1	0,6	8,6	0,1	-	0,3	-	0,2	0,5
Acinetobacter Otu040	-	-	-	2,9	-	-	-	12,9	-	-	-	0,03	-	-	-	2,0

The phylogenetic distribution of 16S rRNA phylotypes was strikingly different in reactors compared to gut inocula. Surprisingly, the sequences related to *Spirochaetes*, *Fibrobacteres* and TG3, highly abundant in termite gut samples, almost disappeared at the end of the fermentation (Figure 3). Sequences related to *Bacteroidetes* were highly abundant in reactors inoculated with *N. lujae* guts as well as in one of the *M. parvus* and *N. ephratae* reactors. This phylum was represented by two OTUs belonging to *Dysgonomas* genus and one

to *Bacteroides*. In the reactors, *Proteobacteria* related OTUs belonged to families of *Gammaproteobacteria*, including members related to *Pseudomonas*, *Escherichia*, *Acinetobacter* and unclassified genus. *Firmicutes* related OTUs were constituted by *Clostridia*, mainly related to *Lachnospiraceae* and *Ruminococcaceae* orders. *Clostridium termitidis* was found in all reactor samples at high frequency, representing in average 16% of the sequences, whereas this species represented less than 0.02% of sequences in the gut inocula.

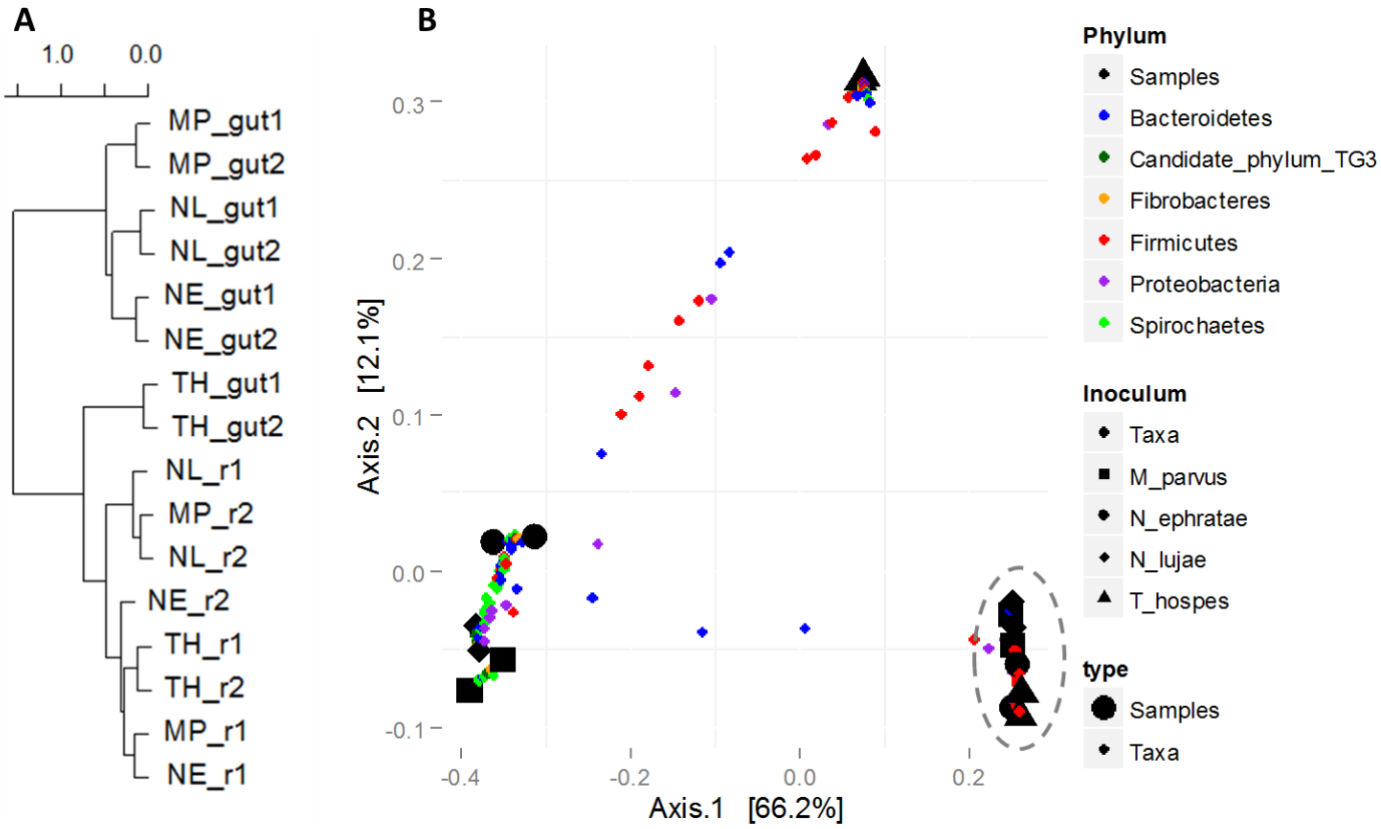


Fig.5. weighted-Unifrac distances between samples were clustered (A) or plotted using PCoA (B). Black dots correspond to samples, their shapes refers to inoculum. Small dots are colored according to their phylum, and correspond to the projection of OTUs in the samples space, so their distance to samples reflects their specificity (minor phyla are not shown). Dash lines encircle reactor samples.

The enrichment in *Firmicutes*, *Proteobacteria* and *Bacteroidetes* observed in the reactors was at the expenses of the three dominant phyla observed in the termite guts. Bacteria that were able to grow in the conditions imposed in the reactors represented less than 5% in the initial gut community, and in some cases they represented less than 0.01% (Table).

II.4. Discussion

The goal of this study was to determine the abilities of termite gut microflora to degrade raw wheat straw as sole carbon source in controlled conditions imposed in bioreactors. This study clearly demonstrate for the first time that termite gut flora from *Nasutitermes ephratae, N. lujae, Microcerotermes parvus* and *Termes hospes* were able to degrade unpretreated lignocellulose substrate and produce carboxylates under the incubation conditions applied. For all termite microbiota studied, wheat straw degradation varied between 32 to 45% after twenty days of incubation. The best lignocellulose conversion was obtained in reactors inoculated with *N. ephrate* guts which displayed high xylanase and cellulase activities, produced up to 5.6±0.3 $g.L^{-1}$ acetate and reached 45±4 % (w/w) wheat straw degradation. Such degradation levels were high compared to similar experiments performed with larger amounts of cow rumen inoculum (Lazuka et al., 2015). *N. lujae* inoculated reactors displayed high wheat straw degradation potential (40±0.5%) but a strong variability on carboxylate production was observed in the replicate reactors: one of them reached 5.3 $g.L^{-1}$ of carboxylate production, presenting a similar stoichiometric conversion than *N. ephrate*, while in the second reactor only a level of 3.4 g $VFA.L^{-1}$ was obtained. A similar behavior, but with lower degradation (32±3%) and VFA accumulation (3.5 ±0.6 $g.L^{-1}$) levels, was observed for reactors inoculated with *M. parvus* guts. In contrast, for the reactors inoculated with *T. hospes* guts, both replicates followed a similar wheat straw degradation kinetics reaching a maximum degradation of 32±3%, accumulating 3.5±0.1 $g.L^{-1}$ of VFA. It has to be noticed that guts from the different termite species were collected from the termite nests at three month interval (Material and Methods section). In the case of *N. lujae*, the high variability observed in wheat straw degradation could be explained by the different inoculum size obtained at each collecting time; indeed, the inocula of the two replicate reactors displayed two-fold difference on 16S rRNA gene copies (Table 1). Nevertheless, it was not the case of *M. parvus* where the abundance on total bacteria in the inocula (estimated as 16S rRNA gene copies) presented only minor differences (15%). Moreover, the community composition of the two inocula of *M. parvus* did not present major changes. It seems thus that other unexplained factors, such as the specific physiological state of the gut flora at the time of the gut's withdraw, could have affected the replicability of the experiments.

In order to explain the lignocellulose degradation performance observed in the termite-gut inoculated reactors, two major enzymatic activities involved on this process were measured in the second series of reactor experiments (reactor 2). All the reactors, irrespective

of the inocula, displayed mainly cell-bound cellulase and xylanase activities whereas extracellular activities were more variable. Reactor inoculated with *T. hospes* guts displayed the lower xylanase and cellulase activities compared to reactors inoculated with guts from other termite species. Such results seem in accordance with the lower lignocellulose degradation capacity observed for the microbiota of *T. hospes*. In reactors inoculated with *N. ephrate, N. lujae* and *M. parvus* guts similar levels of cell-bound cellulase and xylanase activities were measured while some variations were observed for the extracellular enzyme activities. Among them, those inoculated with *N. ephrate* and *N. lujae* guts (r2 for both species) displayed similar wheat straw degradation capacities; they also displayed the highest extracellular activities for both cellulase and xylanase. It seems that enzymes activities could explain the highest degradation levels observed in these reactors. However, the reactor inoculated with *M. parvus* flora displayed similar enzyme activity levels than *N. ephrate* and *N. lujae* reactors, but a lower degradation capacity; indeed, its degradation could be compared to that observed in *T. hospes* reactors whereas its enzymatic activity was two-fold higher. *M. parvus* inoculated reactor only differ from the *Nasutitermes* ones on its lower extracellular CMCase activity. It is possible that such lower cellulase activity have impacted the overall wheat straw degradation efficiency. Previous studies comparing enzyme activity levels in switchgrass and corn stover degradation experiments showed that 34 % degradation was reached with about 58UA xylanase and cellulase while less than 6UA enabled 23% biomass removal (Reddy et al., 2013). It is known that different enzymes families, displaying distinct substrate affinities, enzyme complementarities and synergies, participate to biomass deconstruction (Kumar et al., 2008; Wei et al., 2009). It could thus be expected that wheat straw degradation would vary in function of the specific make-up of enzymes present in each reactor. Further studies on the characterization of the specific hydrolytic enzymes present in the different reactors could explain the differences degradation performances observed and shed light on the mechanisms used for biomass deconstruction.

A second aim of this study was to characterize the termite gut microbial communities and identify the communities involved on wheat straw degradation selected in controlled bioreactors. With this aim, the bioreactors were inoculated with a low inoculum to substrate ratio, representing less than 5% w/w of the initial wheat straw mass. To avoid competition with wheat straw endogenous bacteria, the wheat straw was sterilized. The analysis of sequencing data showed that species abundance in the initial termite gut communities presented important differences.

Among the termite species studied, *M. parvus*, *N. ephrate* and *N. lujae* gut microbiota where dominated by *Spirochaetes* while *T. hospes* was the richest in *Firmicutes* related OTUs. Members of the *Fibrobacteres* phylum were only detected in *M. parvus* and *N. ephrate* guts. These results are consistent with previous descriptions of termite gut communities (Thongaram et al., 2005)(Hongoh et al., 2006). Whereas in the termite phylogeny *Nasutitermes* genus is closer to *Termes* genus than to *Microcerotermes* (Supplementary Information S2), *T. hospes* gut community was clearly the more distant group irrespective of the method used for distance estimation. As *Termes* is the only humus-feeding termite species while the three others are wood-feeding species, our results confirm that termite gut microbiota are mainly shaped by the host diet rather than host phylogeny, as previously proposed by Mikaelyan et al. (Mikaelyan et al., 2015a). Moreover, *T. hospes* gut community was very close to the *T. comis* already described (Thongaram et al., 2005); it seems thus that bacterial profiles are robust inside the *Termes* genus. For the wood-feeding termites, *N. ephratae* and *M. parvus* gut communities were very similar to *N. takasagoensis* microbiome (Hongoh et al., 2006), suggesting a great gut microbiome conservation inside this genus. Surprisingly, weighted-Unifrac distance of gut microbial communities from *M. parvus, N. lujae* and *N. ephrateae* were almost equally distant, but not with Unifrac distance. That means that these termite gut communities shared numerous OTUs but their relative abundances were prone to variation between species of a same genus. Indeed, the large weighted-Unifrac distance observed between communities of *N. ephratae* and *N. lujae* gut resulted from the differences in abundance of *Fibrobacteres* compared to the two other termite gut communities. The presence of TG3-related OTUs in all termite guts confirmed that this hypothetic new phylum is largely widespread and dominant in termites gut communities. TG3-related OTUs were surprisingly well grouped by host (Supplementary Information S3), suggesting co-evolution and host-adaptation processes as proposed by Hongoh et al. (Hongoh et al., 2005). TG3 members have been described as associated to fibers in termite hindgut, and they are considered as putative lignocellulolytic bacteria (Mikaelyan et al., 2014). The two other major taxa observed in termite guts were *Spirochaetes* and *Fibrobacteres* which are not well characterized. However, they have been described as cellulose degraders in ruminant animals (Kobayashi et al., 2008) and fiber-associated lignocellulose degraders in higher termites (Mikaelyan et al., 2014; Warnecke et al., 2007). *Spirochaetaceae* have been identified as vital for termites (Eutick et al., 1978) where they are known to produce acetate from H_2 plus CO_2 (Brauman et al., 1992); it could be an interesting metabolic pathway for

carboxylate production. However, despite their large initial abundance, none of these phyla grew in the wheat straw degradation reactors, disappearing after twenty days of incubation.

Indeed, after the incubation on wheat straw degradation reactors, the diversity analysis bases on Shanon's and Simpson's reciprocal indices, showed a strong decrease on diversity between the initial gut communities and the reactor communities. For all termite gut inocula, the conditions imposed in wheat straw bioreactors selected mainly *Firmicutes*, *Bacteroidetes* and *Proteobacteria* related OTUs. Particularly, reactors inoculated with *N. lujae* guts displayed very reproducible community make-up, constituted by these three phyla; for reactors inoculated with *T. hospes* guts, the community composition was also very stable, displaying a very low abundance on *Bacteroidetes* related OTUs for both replicate reactors. The final community composition of reactors inoculated with *N. ephrate* and *M. parvus* guts displayed stronger variability, in one case a large abundance of *Firmicutes* with low abundance of *Bacteroidetes* was observed, whereas in the second reactor the community was constituted by the three dominant phyla *Firmicutes*, *Bacteroidetes* and *Proteobacteria*. The members related to *Proteobacteria* were affiliated to *Enterobacteriaceae, Pseudomonas, Acinetobacter, Stenotrophomonas* and *Rhodocyclales*. These groups are not known for their ability to degrade lignocellulose but they are able to ferment carbohydrates derived from lignocellulose degradation (Imhoff, 2005). Surprisingly, members related to *Stenotrophomonas*, *Pseudomonas* and *Acinetobacter*, groups known for displaying an aerobic metabolism, were observed in the lignocellulose reactors; their prevalence under the strict anaerobic conditions imposed in the reactors and in association with strict anaerobes such as *Clostridium, Bacteroides or Acetanaerobacterium* remained to be explained. In all the lignocellulose reactors, OTUs belonging to *Clostridia* (*Firmicutes*), particularly *Clostridium termitidis* OTU_{1}, as well as OTUs related to *Lachnospiraceae* ($OTU_{3, 4, 37, 43}$) and *Ruminococcaceae* ($OTU_{19, 23, 24, 53}$) were strongly enriched. *C. termitidis* was present in all reactors at an average of 16%. This species was firstly described in *N. lujae* gut (Hethener et al., 1992) and it is a cellulose degrader able to use various sugars including xylose as carbon source. The exact role *Ruminococcaceae* and *Lachnospiraceae* in lignocellulose degradation remains unknown but they have been described in termite environments, mainly in association to fungus-feeding or humus and soil feeder termites (Mikaelyan et al., 2015a). Furthermore, it is known that clostridial genomes are enriched in xylanases genes and present cohesins and dockerins that are key components of cellulosomes; it is thus possible that Clostridiales are involved in lignocellulose digestion (He et al., 2013). Members of the *Firmicutes* phylum, which were enriched in all the reactors, appeared thus as the main candidates for wheat straw degradation

under the studied conditions. OTUs related to *Bacteroidetes* were also enriched in some of the reactors, particularly *Dysgonomonas* ($OTU_{2, 15}$) and *Bacteroides* (OTU_{18}) related OTUs were present in reactors inoculated with *N. lujae*, *M. parvus* and *N. ephrate* guts. *Dysgonomonas* genus is known for their lignocellulolytic potential (Sun et al., 2015) and has been identified as a putative cellulose degrader in termite gut (Yang et al., 2014); this genus is also able to degrade cellobiose and glucose (Hofstad et al., 2000). *Bacteroides* has been often found in anaerobic mammalian digestive tracts and it has been described as xylanolytic (Tomomi Nishiyama, 2009).

The final community composition of wheat straw degradation reactors was clearly distinct to the initial termite gut communities (Figure 4). Diversity analysis showed that OTUs present in termite guts were host-specific, whereas in final reactor communities most of the enriched OTUs were shared between the different reactors, showing a convergence of the selected bacteria under the conditions imposed. Moreover, the reactor community make-up was closer to lignocellulolytic communities present in rumen or soil than to termite gut communities (He et al., 2013). Factors such as pH, nutrient and O2 requirements, host specific signals or inter- taxa dependency are some of the possible explanations of the loss of these phyla during reactor incubation.

It is difficult to compare the wheat straw degradation capacity of termite guts microbiota studied here with that reported in previous studies. Indeed, lignocellulose degradation have been frequently assessed using enriched microbial communities from compost, forest soils, mangrove sediments (Feng et al., 2011; Reddy et al., 2011; Yan et al., 2012). Such studies reported degradation levels that vary strongly in function of the lignocellulosic substrate and culture conditions. For example, Feng et. al. (2011) reported 51% degradation of corn stover powder (mesh 40 or 375µm) and 44% degradation of corn stover pretreated by steam-explosion using an enriched consortium from woodland soil incubated at 40°C. In another study, 49% degradation of alkali-pretreated rice straw was reported, after 7 days of incubation, using an enriched microbial consortium issued from an anaerobic digester treating cow manure (Yan et al., 2012). However, to our knowledge, no previous study concerns the lignocellulose degradation by termite gut microbiota implemented in bioreactors. Degradation efficiencies reported here are comparable to those observed in previous experiments despite the different experimental conditions used. In this work, the best wheat straw degradation (45%) was observed in reactors inoculated with

Nasutitermes flora. Nevertheless, it could be possible to improve such degradation levels, as well as that of the other termite microbiomes, by enriching the lignocellulolytic community and reconsidering the conditions imposed in the bioreactors. In the conditions applied here, a slightly acid pH was chosen in order to inhibit methanogenesis and favor carboxylate accumulation. However, most of termites have an alkaline gut. Indeed, lignocellulose degradation in termites occurs is the P3 compartment of the hindgut, which harbors most of the microbial flora. The pH conditions prevailing in this compartment are not well documented. Data is only available for some soil-feeding higher termites such as genus *Cubitermes* which present an extremely alkaline gut (Brune and Kühl, 1996) while the P3 segment in *Nasutitermes* genus present a neutral or even slightly acid pH (Köhler et al., 2012). This particular pH, close to that used in this study may explain why *Nasutitermes* flora performed better that other termite microbiota under the studied conditions. Moreover, the microbial community acting on lignocellulose deconstruction could also be enriched by sub-culturing approaches in order to improve the degradation efficiency (Reddy et al., 2012; Lazuka et al., 2015).

II.5. Conclusion

This work demonstrate that gut microflora from the termites were able to degrade lignocellulose in reactor conditions; these results provide motivation to further identify the key bacteria and enzymes involved in biomass deconstruction in such termite gut microbiomes and optimize their wheat straw degradation and carboxylate production in bioconversion processes.

II.6. Experimental procedures

II.6.1 Lignocellulose substrate and termite gut inocula

Wheat straw from the winter wheat variety Koreli was collected at an experimental farm (INRA, Boissy-le-Repos, France) in August 2011. After harvest, the straw was milled to 2 mm and stored at room temperature (20-25°C).

Four different species of higher termites were selected as source of inocula for this study: *Microcerotermes parvus, Termes hospes, Nasutitermes ephratae* and one non-described species closely related to *Nasutitermes lujae.* They were selected based on to their intestinal flora composition, dominated by bacteria, and their intestinal pH, slightly acidic or near to the neutrality (Köhler et al., 2012). *M. parvus, N. ephratae* and to *N. lujae* are wood-

feeding termites while *T. hospes* is humus-feeding species (Eggleton et al., 1995). Other factors such as knowledge on their lignocellulose degradation capacities, acetate production, and availability at the Institut de Recherche pour le Developpement (IRD, Bondy, France) were also considered.

Termite colonies were maintained in a climate-controlled room (27°C, 60% relative humidity) at the IRD. One thousand worker termites per species were randomly collected from their nest at tree month interval (n =500 guts per species at each time). After cold anesthesia of termites, the whole gut was removed by dissection on ice with sterile scissors and forceps; guts were immediately placed into a physiological saline solution maintained on ice. Each 500 gut's collection, gut samples were frozen at -20°C. For DNA extraction, 20 guts per species were separately collected (in duplicate) and stored. Gut samples were transported to our laboratory on dry ice and stored at -80°C until use. Groups of ten guts were weighted to estimate the inoculum biomass.

II.6.2 Anaerobic reactors

To assess the lignocellulose degradation capacity of the different gut microbiota, two replicate anaerobic reactors (Applikon MiniBio 500) were conducted for each termite species. After centrifugation (7197g, 10min, 4°C) and elimination of the saline media, 500 guts were used to inoculate 400mL mineral media (MM). MM contained, per liter of distilled water: KH_2PO_4, 0.45g, K_2HPO_4, 0.45g; NH_4Cl, 0.4g; NaCl, 0.9g; $MgCl_2.6H_2O$, 0.15g; $CaCl_2.2H_2O$, 0.09g. It was supplemented with 250µL of V7 vitamin solution (Pfennig and Trüper, 1992), and 1mL trace elements solution, containing per liter of distilled water: H_3BO_3, 300 mg; $FeSO_4.7H_2O$, 1.1 g; $CoCl_2.6H_2O$, 190 mg; $MnCl_2.4H_2O$, 50 mg; $ZnCl_2$, 42 mg; $NiCl_2.6H_2O$, 24 mg; $NaMoO_4.2H_2O$, 18 mg; $CuCl_2.2H_2O$, 2 mg; sterilized by filtration (0.2 µm). The milled wheat straw (2mm) was sterilized by autoclaving (120°C, 20 min and 1.2 bars) and added to the medium ($20g.L^{-1}$) as sole carbon source. Reactors were conducted on strict anaerobic conditions obtained by flushing the reactor with nitrogen after inoculation. Temperature was set to 35°C and pH was controlled to 6.15 by adding a 2M NaOH solution under constant steering (400 rpm). During the incubation, if methane production was detected, a solution of *2-bromoethanesulfonate* (BES), a methanogenesis inhibitor, was spiked until a maximum concentration of 10 mM. As a control, a non-inoculated reactor was also conducted in the same conditions described above. Samples were collected every two days during 20 days of incubation to characterize the LC degradation capacities of the different termite gut inocula.

II.6.3 Chemical analysis

Gas production was monitored by pressure measurement and gas composition was analyzed using a gas chromatograph (GC) HP 5890 equipped with a conductivity detector and a HAYSEP D column (molecular sieve of 5 Å). Argon was used as the carrier gas at a flowrate of 100mL.min-1. Injector, oven, and detector temperatures were 100, 60, and 140 °C, respectively.

Volatile Fatty Acids (VFA) composition and concentration were determined in liquid phase of samples regularly withdrawn from the reactor, using a Varian 3900 gas chromatograph equipped with a flame ionization detector and CP-Wax 58 (FFAP) CB column (length: 25m, inside diameter: 0.53 mm).

Lignocellulose concentration was measured at the beginning and at end of the incubation (20 days) by measuring the total (TS) and volatile (VS) solids. TS were determined using 10 mL samples that were first centrifuged (7197 x g, 10 min), rinsed twice with distilled water and dried 24h at 105°C. The mineral fraction (MF) was estimated by mineralization of the samples at 500°C for 2h, and VS were determined from the difference between TS and MF. Wheat straw degradation is reported as percentage (%, w/w) related to the initial wheat straw mass.

The initial and final wheat straw biochemical composition was determined using the sulfuric acid hydrolysis method described by de Souza et al. (2013). Complete LC hydrolysis was achieved in two steps. First, a sample (40 mg) was prehydrolyzed by incubation at 60 min at 30°C in 500 µL concentrated H2SO4 (72% w/w). Second, the sample was diluted by the addition of H_2O to reach a final acid concentration of 10% (w/w) and then incubated at 100°C for 90 min. Quantification of monomeric sugars was achieved using high performance liquid chromatography (HPLC) on a Ultimate 3000 Dionex separation system equipped with a BioRad Aminex HPX 87H affinity column and a refractive index detector (Thermo Scientific). The separation of product species was performed in H_2SO_4 (5 mM) at 40°C and a flow rate of 0.3mL.min^{-1}.

II.6.4 Enzyme activity assay

Enzyme activity measurements were performed as previously described by Lazuka et al. (Lazuka et al., 2015). Briefly, triplicate reactor samples (5 mL) were removed at the end of the experiment (20 d) and centrifuged at 7197 x g for 10 min at 4°C. Supernatant was

considered to contains extracellular enzymes, while the pellet represents cell-bound enzymes. Xylanase and endoglucanase (CMCase) activities were measured using 1% w/v xylan beechwood (Sigma) and 1% w/v carboxymethyl cellulose (CMC) (Sigma). Activities were estimated by the DNS method as previously described.

II.6.5 16 rRNA gene copy number and diversity analysis

16S rRNA gene copy number and bacterial diversity were analyzed on the initial termite guts and at the end of the bioreactor incubation. 1.5mL samples were collected and centrifuged at 13 000 x g, 5 min, 4°C. After removing the supernatant, the pellet was snap frozen in liquid nitrogen and stored at -80°C until nucleic acid extraction. Total DNA/RNA were extracted from these samples using a PowerMicrobiome RNA Isolation kit (MoBio Laboratories Inc. Carlsbad) following the manufacturer's instructions but omitting the final DNAse steps. Cell lysis was carried out with a Fast Prep (MP Biomedicals) (2 x 30s at $4ms^{-1}$). DNA and RNA were separated and purified using an AllPrep DNA/RNA Mini Kit (Qiagen) according to the manufacturer's instructions. DNA integrity and purity was checked by agarose gel (1%) electrophoresis. DNA concentration was measured by NanoDrop 1000 spectrophotometer (Thermo Scientific), measuring absorbance at 260 and 280 nm.

16S rRNA copy number was determined by qPCR using a Realplex Mastercycler (Eppendorf, Montesson, France); assays were carried out in triplicate for each sample using 96-well real-time PCR plates (Eppendorf). The qPCR was performed in 25µL containing 12.5µL Master Mix (Invitrogen, Eugen, USA), using primers BAC3388 and BAC805R (250nM of each primer), the TaqMan probe BAC516F (100nM) and DNA template ranging from 10 ng to 100 ng as previously described (Yu et al., 2005). The real-time PCR thermocycling was set as follows: 95°C for 20 sec and 40 cycles at 95°C for 15 sec and 60°C for 1 min. A negative control without DNA template was subjected to the same procedure to exclude any possible contamination. One standard curve was generated for each assay by using 10-fold dilutions of pEX-A plasmids (Eurofins MWG Operon) containing the targeted gene sequence. Three different dilutions of each sample were amplified and the initial concentrations were calculated from reactions displaying no PCR inhibition.

Microbial diversity was analyzed using MiSeq Illumina sequencing performed by the GenoToul Genomics and Transcriptomics facility (GeT, Auzeville, France). V3-V4 hypervariable region of the 16S rRNA gene was amplified from genomic DNA samples with

the bacterial primers 343F and 784R, modified to add adaptors during the second PCR : 343F= 5'-CTT TCC CTA CAC GAC GCT CTT CCG ATC TAC GGR AGG CAG CAG-3' and 784R= 5'-GGA GTT CAG ACG TGT GCT CTT CCG ATC TTA CCA GGG TAT CTA ATC CT-3'. The first PCR was performed in 50µl reaction mixture containing 1X PCR buffer, 2.5U MTP Taq DNA Polymerase (Sigma), 0.2mM of each dNTP, 0.5mM of each primer and 2ng of extracted DNA. After 30 amplification cycles of 94°C-65°C-70°C, one minute each step, amplicons were purified using magnetic beads and quantified by NanoDrop 1000 spectrophotometer. A second PCR was performed at the GeT platform to add sequencing adapters and a unique index for each sample (details in Supplementary data). The PCR products were purified again with magnetic beads. Amplicon quality was then checked with High Sensivity DNA Analysis Kits (Agilent) and a BioAnalyzer 2100; DNA concentration was measured by NanoDrop 1000 spectrophotometer. An equimolar pool was then prepared and loaded on a MiSeq Illumina cartridge, using reagent kit v2. MiSeq v2 reagents enabled paired 250-bp reads.

Data were demultiplexed at the GeT platform and pair-ends reads were joined with Flash v1.2.6 (Magoč and Salzberg, 2011), using an overlap bigger than 110bp with a maximum ratio of 0.1 mismatches, which generated high quality full-length reads of the V3 and V4 regions. All the fastq files were then merged into a unique fasta file, treated with Mothur v1.33.1 (Schloss et al., 2009). Sequences presenting a primer mismatch or that did not present the expected length (380 to 460bp) were discharged. To reduce the computational costs, sequences were dereplicated and unique sequences were aligned with SILVA database and only the sequences aligning in the expected V3-V4 region were further analyzed. Sequences that presented less than 5 differences with a more abundant one were considered as sequencing errors and were merged with the most abundant one. Chimeras were detected and removed with Uchime (Edgar et al., 2011) using a self-reference and default parameters on each sample groups. Sequences were then clustered at 3% distance and taxonomic affiliations were obtained with Wang method and a fusion of LTP v115 (Yarza et al., 2008) and DictDB (Mikaelyan et al., 2015b) databases . Rare OTUs, containing less than 20 sequences across all samples were removed, and a random subsampling normalizes all samples to 15k sequences. Abundance tables with affiliation and rarefaction curves were generated with Mothur.

Abundance tables, taxonomy files and phylogenetic trees were manually imported into R (v3.0.3) package phyloseq v1.6.1 (McMurdie and Holmes, 2013). Unifrac and weighted-Unifrac distances were calculated with Phyloseq and clustered with Hclust. Heatmaps and

PCoA plots were generated using Phyloseq. ClustalOWS alignment and calculation of neighbor-joining or average distance trees were performed with Jalview v2.8.2 (Waterhouse et al., 2009).

II.7. Acknowledgments

This research was supported by the French National Institure for Agronomical Reaseach (INRA) and the Region Languedoc-Roussillon Midi-Pyrénées and was in collaboration with the Institute for Reasearch and Development (IRD). The authors thank the Genomics and Transcripmics (GeT) INRA platform for their help with sequencing. Alain Robert and Isabel Monteiro are acknowledge for their help with termite nests, and Gunnar Oelker for its assistance with experiments and fermenters.

II.8. References

Agler, M.T., Wrenn, B.A., Zinder, S.H., and Angenent, L.T. (2011). Waste to bioproduct conversion with undefined mixed cultures: the carboxylate platform. Trends Biotechnol. *29*, 70–78.

Barakat, A., de Vries, H., and Rouau, X. (2013). Dry fractionation process as an important step in current and future lignocellulose biorefineries: A review. Bioresour. Technol. *134*, 362–373.

Brauman, A., Kane, M.D., Labat, M., and Breznak, J.A. (1992). Genesis of acetate and methane by gut bacteria of nutritionally diverse termites. Science *257*, 1384–1387.

Breznak, J.A., and Brune, A. (1994). Role of Microorganisms in the Digestion of Lignocellulose by Termites. Annu. Rev. Entomol. *39*, 453–487.

Brune, A. (2014). Symbiotic digestion of lignocellulose in termite guts. Nat. Rev. Microbiol. *12*, 168–180.

Brune, A., and Kühl, M. (1996). pH profiles of the extremely alkaline hindguts of soil-feeding termites (Isoptera: Termitidae) determined with microelectrodes. J. Insect Physiol. *42*, 1121–1127.

Chandel, A.K., and Singh, O.V. (2011). Weedy lignocellulosic feedstock and microbial metabolic engineering: advancing the generation of "Biofuel." Appl. Microbiol. Biotechnol. *89*, 1289–1303.

Cope, J.L., Hammett, A.J.M., Kolomiets, E.A., Forrest, A.K., Golub, K.W., Hollister, E.B., DeWitt, T.J., Gentry, T.J., Holtzapple, M.T., and Wilkinson, H.H. (2014). Evaluating the performance of carboxylate platform fermentations across diverse inocula originating as sediments from extreme environments. Bioresour. Technol. *155*, 388–394.

Deacon, J. (2005). Fungal biology (Oxford, UK: Blackwell Publishing).

DeAngelis, K.M., Fortney, J.L., Borglin, S., Silver, W.L., Simmons, B.A., and Hazen, T.C. (2012). Anaerobic decomposition of switchgrass by tropical soil-derived feedstock-adapted consortia. mBio *3*.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. Bioinformatics *27*, 2194–2200.

Eggleton, P., Bignell, D.E., Sands, W.A., Waite, B., Wood, T.G., and Lawton, J.H. (1995). The Species Richness of Termites (Isoptera) Under Differing Levels of Forest Disturbance in the Mbalmayo Forest Reserve, Southern Cameroon. J. Trop. Ecol. *11*, 85–98.

Eutick, M.L., Veivers, P., O'Brien, R.W., and Slaytor, M. (1978). Dependence of the higher termite, Nasutitermes exitiosus and the lower termite, Coptotermes lacteus on their gut flora. J. Insect Physiol. *24*, 363–368.

Feng, Y., Yu, Y., Wang, X., Qu, Y., Li, D., He, W., and Kim, B.H. (2011). Degradation of raw corn stover powder (RCSP) by an enriched microbial consortium and its community structure. Bioresour. Technol. *102*, 742–747.

Guo, P., Zhu, W., Wang, H., Lü, Y., Wang, X., Zheng, D., and Cui, Z. (2010). Functional characteristics and diversity of a novel lignocelluloses degrading composite microbial system with high xylanase activity. J. Microbiol. Biotechnol. *20*, 254–264.

He, S., Ivanova, N., Kirton, E., Allgaier, M., Bergin, C., Scheffrahn, R.H., Kyrpides, N.C., Warnecke, F., Tringe, S.G., and Hugenholtz, P. (2013). Comparative metagenomic and metatranscriptomic analysis of hindgut paunch microbiota in wood- and dung-feeding higher termites. PloS One *8*, e61126.

Hethener, P., Brauman, A., and Garcia, J.-L. (1992). Clostridium termitidis sp. nov., a Cellulolytic Bacterium from the Gut of the Wood-feeding Termite, Nasutitermes lujae. Syst. Appl. Microbiol. *15*, 52–58.

Hobson, P.N. (1998). The rumen microbial ecosystem. (New York, NY USA: Elsevier Sci. Publisher Ltd.).

Hofstad, T., Olsen, I., Eribe, E.R., Falsen, E., Collins, M.D., and Lawson, P.A. (2000). Dysgonomonas gen. nov. to accommodate Dysgonomonas gadei sp. nov., an organism isolated from a human gall bladder, and Dysgonomonas capnocytophagoides (formerly CDC group DF-3). Int. J. Syst. Evol. Microbiol. *50 Pt 6*, 2189–2195.

Hollister, E.B., Hammett, A.M., Holtzapple, M.T., Gentry, T.J., and Wilkinson, H.H. (2011). Microbial community composition and dynamics in a semi-industrial-scale facility operating under the MixAlco^{TM} bioconversion platform. J. Appl. Microbiol. *110*, 587–596.

Holtzapple, M.T., Davison, R.R., Ross, M.K., Albrett-Lee, S., Nagwani, M., Lee, C.M., Lee, C., Adelson, S., Kaar, W., Gaskin, D., et al. (1999). Biomass conversion to mixed alcohol fuels using the MixAlco process. Appl. Biochem. Biotechnol. *77–79*, 609–631.

Hongoh, Y., Deevong, P., Inoue, T., Moriya, S., Trakulnaleamsai, S., Ohkuma, M., Vongkaluang, C., Noparatnaraporn, N., and Kudo, T. (2005). Intra- and Interspecific Comparisons of Bacterial Diversity and Community Structure Support Coevolution of Gut Microbiota and Termite Host. Appl. Environ. Microbiol. *71*, 6590–6599.

Hongoh, Y., Deevong, P., Hattori, S., Inoue, T., Noda, S., Noparatnaraporn, N., Kudo, T., and Ohkuma, M. (2006). Phylogenetic diversity, localization, and cell morphologies of members of the candidate phylum TG3 and a subphylum in the phylum Fibrobacteres, recently discovered bacterial groups dominant in termite guts. Appl. Environ. Microbiol. *72*, 6780–6788.

Imhoff, J.F. (2005). "Enterobacteriales." In Bergey's Manual® of Systematic Bacteriology, D.J. Brenner, N.R. Krieg, J.T. Staley, G.M.G. Sc.D, D.R. Boone, P.D. Vos, M. Goodfellow, F.A. Rainey, and K.-H. Schleifer, eds. (Springer US), pp. 587–850.

Kleerebezem, R., and van Loosdrecht, M.C. (2007). Mixed culture biotechnology for bioenergy production. Curr. Opin. Biotechnol. *18*, 207–212.

Kobayashi, Y., Shinkai, T., and Koike, S. (2008). Ecological and physiological characterization shows that Fibrobacter succinogenes is important in rumen fiber digestion — Review. Folia Microbiol. (Praha) *53*, 195–200.

Köhler, T., Dietrich, C., Scheffrahn, R.H., and Brune, A. (2012). High-resolution analysis of gut environment and bacterial microbiota reveals functional compartmentation of the gut in wood-feeding higher termites (Nasutitermes spp.). Appl. Environ. Microbiol. *78*, 4691–4701.

König, H., Li, L., and Fröhlich, J. (2013). The cellulolytic system of the termite gut. Appl. Microbiol. Biotechnol. *97*, 7943–7962.

Kumar, R., Singh, S., and Singh, O.V. (2008). Bioconversion of lignocellulosic biomass: biochemical and molecular perspectives. J. Ind. Microbiol. Biotechnol. *35*, 377–391.

Lazuka, A., Auer, L., Bozonnet, S., Morgavi, D.P., O'Donohue, M., and Hernandez-Raquet, G. (2015). Efficient anaerobic transformation of raw wheat straw by a robust cow rumen-derived microbial consortium. Bioresour. Technol. *196*, 241–249.

Lo, N., Tokuda, G., and Watanabe, H. (2010). Evolution and Function of Endogenous Termite Cellulases. In Biology of Termites: A Modern Synthesis, D.E. Bignell, Y. Roisin, and N. Lo, eds. (Springer Netherlands), pp. 51–67.

Lynd, L.R., Weimer, P.J., van Zyl, W.H., and Pretorius, I.S. (2002). Microbial cellulose utilization: fundamentals and biotechnology. Microbiol. Mol. Biol. Rev. MMBR *66*, 506–577, table of contents.

Magoč, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinforma. Oxf. Engl. *27*, 2957–2963.

McMurdie, P.J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PloS One *8*, e61217.

Mikaelyan, A., Strassert, J.F.H., Tokuda, G., and Brune, A. (2014). The fibre-associated cellulolytic bacterial community in the hindgut of wood-feeding higher termites (Nasutitermes spp.). Environ. Microbiol. *16*, 2711–2722.

Mikaelyan, A., Dietrich, C., Köhler, T., Poulsen, M., Sillam-Dussès, D., and Brune, A. (2015a). Diet is the primary determinant of bacterial community structure in the guts of higher termites. Mol. Ecol. *24*, 5284–5295.

Mikaelyan, A., Köhler, T., Lampert, N., Rohland, J., Boga, H., Meuser, K., and Brune, A. (2015b). Classifying the bacterial gut microbiota of termites and cockroaches: A curated phylogenetic reference database (DictDb). Syst. Appl. Microbiol. *38*, 472–482.

Ni, J., and Tokuda, G. (2013). Lignocellulose-degrading enzymes from termites and their symbiotic microbiota. Biotechnol. Adv. *31*, 838–850.

Pfennig, N., and Trüper, H.G. (1992). The Family Chromatiaceae. In The Prokaryotes, A. Balows, H.G. Trüper, M. Dworkin, W. Harder, and K.-H. Schleifer, eds. (Springer New York), pp. 3200–3221.

Reddy, A.P., Allgaier, M., Singer, S.W., Hazen, T.C., Simmons, B.A., Hugenholtz, P., and VanderGheynst, J.S. (2011). Bioenergy feedstock-specific enrichment of microbial populations during high-solids thermophilic deconstruction. Biotechnol. Bioeng. *108*, 2088–2098.

Reddy, A.P., Simmons, C.W., Claypool, J., Jabusch, L., Burd, H., Hadi, M.Z., Simmons, B.A., Singer, S.W., and VanderGheynst, J.S. (2012). Thermophilic enrichment of microbial communities in the presence of the ionic liquid 1-ethyl-3-methylimidazolium acetate. J. Appl. Microbiol. *113*, 1362–1370.

Reddy, A.P., Simmons, C.W., D'haeseleer, P., Khudyakov, J., Burd, H., Hadi, M., Simmons, B.A., Singer, S.W., Thelen, M.P., and Vandergheynst, J.S. (2013). Discovery of microorganisms and enzymes involved in high-solids decomposition of rice straw using metagenomic analyses. PloS One *8*, e77985.

Scharf, M.E. (2015). Omic research in termites: an overview and a roadmap. Front. Genet. *6*, 76.

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., et al. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl. Environ. Microbiol. *75*, 7537–7541.

Slaytor, M. (1992). Cellulose digestion in termites and cockroaches: What role do symbionts play? Comp. Biochem. Physiol. Part B Comp. Biochem. 775–784.

Smant, G., Stokkermans, J.P., Yan, Y., de Boer, J.M., Baum, T.J., Wang, X., Hussey, R.S., Gommers, F.J., Henrissat, B., Davis, E.L., et al. (1998). Endogenous cellulases in animals: isolation of beta-1, 4-endoglucanase genes from two species of plant-parasitic cyst nematodes. Proc. Natl. Acad. Sci. U. S. A. *95*, 4906–4911.

Sun, X., Yang, Y., Zhang, N., Shen, Y., and Ni, J. (2015). Draft Genome Sequence of Dysgonomonas macrotermitis Strain JCM 19375T, Isolated from the Gut of a Termite. Genome Announc. *3*.

Thongaram, T., Hongoh, Y., Kosono, S., Ohkuma, M., Trakulnaleamsai, S., Noparatnaraporn, N., and Kudo, T. (2005). Comparison of bacterial communities in the alkaline gut segment among various species of higher termites. Extrem. Life Extreme Cond. *9*, 229–238.

Tokuda, G., and Watanabe, H. (2007). Hidden cellulases in termites: revision of an old hypothesis. Biol. Lett. *3*, 336–339.

Tomomi Nishiyama, A.U. (2009). Bacteroides graminisolvens sp. nov., a xylanolytic anaerobe isolated from a methanogenic reactor treating cattle waste. Int. J. Syst. Evol. Microbiol. *59*, 1901–1907.

Warnecke, F., Luginbuehl, P., Ivanova, N., Ghassemian, M., Richardson, T.H., Stege, J.T., Cayouette, M., McHardy, A.C., Djordjevic, G., Aboushadi, N., et al. (2007). Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature *450*, 560-U17.

Watanabe, H., Noda, H., Tokuda, G., and Lo, N. (1998). A cellulase gene of termite origin. Nature *394*, 330–331.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189–1191.

Wei, H., Xu, Q., Taylor, L.E., Baker, J.O., Tucker, M.P., and Ding, S.-Y. (2009). Natural paradigms of plant cell wall degradation. Curr. Opin. Biotechnol. *20*, 330–338.

Yan, L., Gao, Y., Wang, Y., Liu, Q., Sun, Z., Fu, B., Wen, X., Cui, Z., and Wang, W. (2012). Diversity of a mesophilic lignocellulolytic microbial consortium which is useful for enhancement of biogas production. Bioresour. Technol. *111*, 49–54.

Yang, Y.-J., Zhang, N., Ji, S.-Q., Lan, X., Zhang, K., Shen, Y.-L., Li, F.-L., and Ni, J.-F. (2014). Dysgonomonas macrotermitis sp. nov., isolated from the hindgut of a fungus-growing termite. Int. J. Syst. Evol. Microbiol. *64*, 2956–2961.

Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K.-H., Ludwig, W., Glöckner, F.O., and Rosselló-Móra, R. (2008). The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst. Appl. Microbiol. *31*, 241–250.

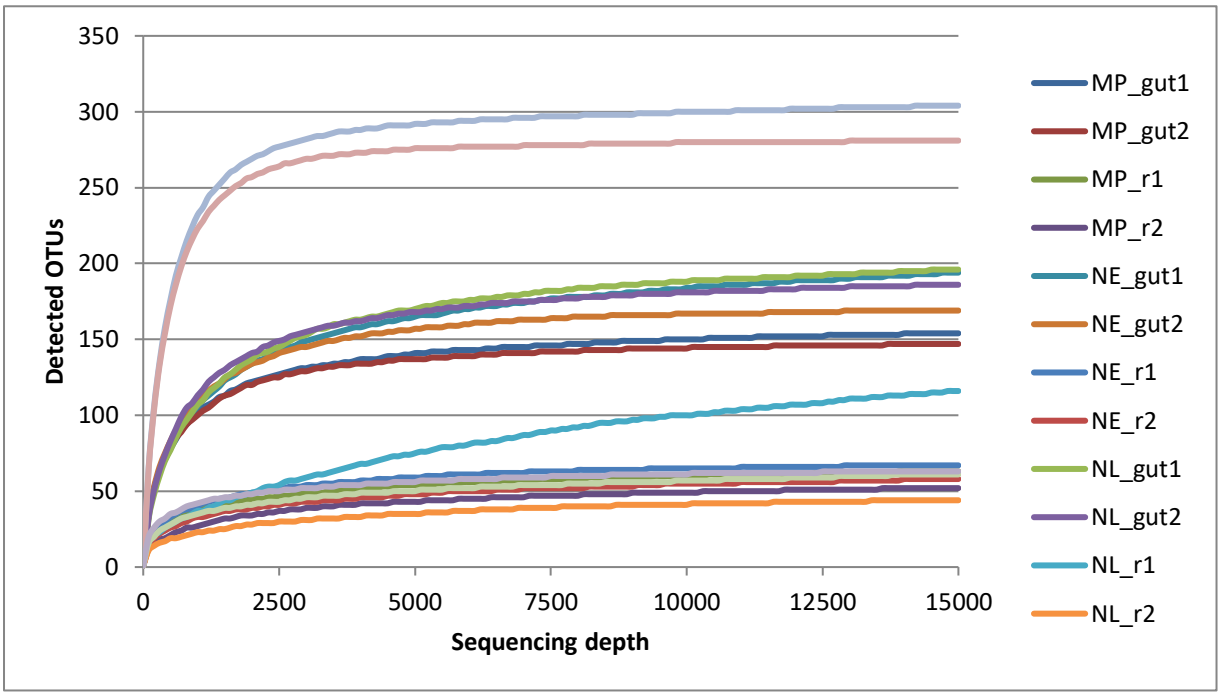
Yu, Y., Lee, C., Kim, J., and Hwang, S. (2005). Group-specific primer and probe sets to detect methanogenic communities using quantitative real-time polymerase chain reaction. Biotechnol. Bioeng. *89*, 670–679.

Yue, Z.-B., Yu, H.-Q., Harada, H., and Li, Y.-Y. (2007). Optimization of anaerobic acidogenesis of an aquatic plant, Canna indica L., by rumen cultures. Water Res. *41*, 2361–2370.

II.9. Supplementary data

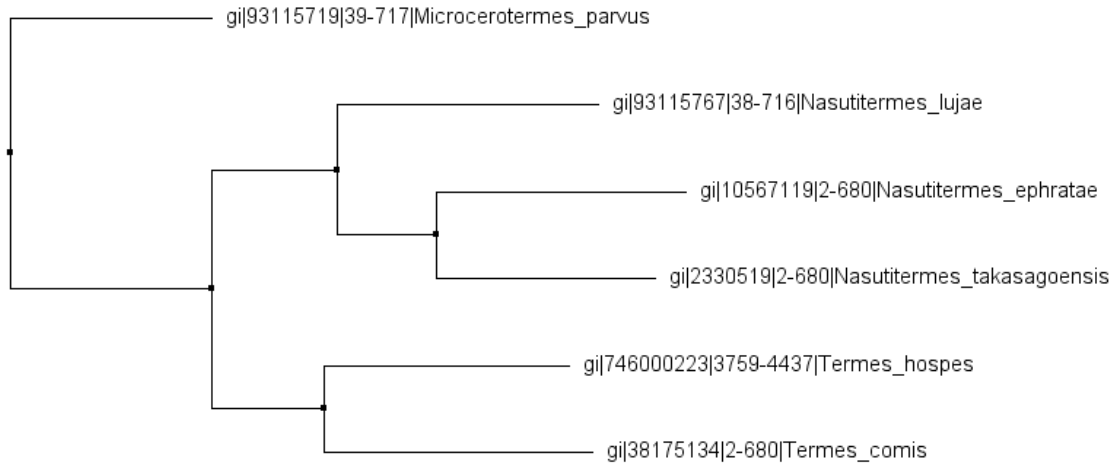
S1. Rarefaction curves of the sequenced points.

Rarefaction curves were generated with the Mothur subroutine rarefaction.single on the randomly subsampled 20k final sequences. For all the samples but *N. lujae* r2, the number of detected OTUs reached a plateau corresponding to its observed richness, indicating that sequencing depth is sufficient to describe the community.

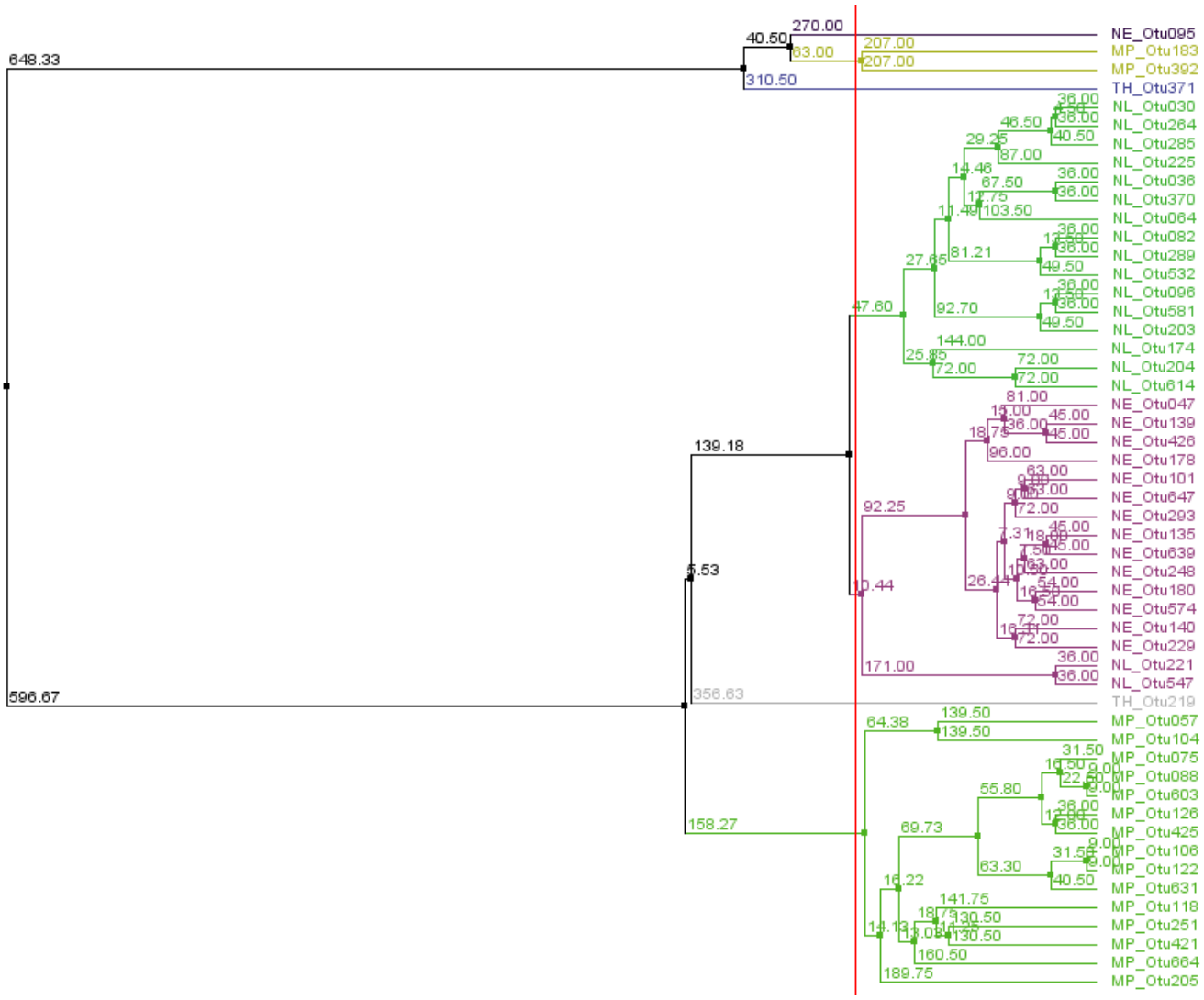


S2. Phylogenetic tree of representative termite species.

Sequences corresponding to the cytochrome oxidase subunit II gene were collected from NCBI. They were aligned using Clustal Omega and use to build a neighbor-joining tree between the four selected species of this study, and two species, *Termes comis* and *Nasutitermes takasagoensis* whose flora were described in other studies and were very close to respectively our *T. hospes* and *N. ephratae*.



S3. Phylogenetic position of the TG3 OTUs



III. Conclusion du chapitre

L'inoculation de fermenteurs par du microbiote intestinal de termite a permis d'atteindre des taux de dégradation de 45% en 15 jours avec *Nasutitermes ephratae*, l'espèce présentant les meilleurs résultats. Cette dégradation, atteinte en un seul cycle de culture est très prometteuse si on la compare à la dégradation obtenue au premier cycle de culture avec du rumen bovin. Celui-ci avait permis d'obtenir seulement 20% de dégradation (mais pour 8 jours seulement, cf chapitre IV), pour atteindre 40% après une procédure d'enrichissement. Malgré sa difficulté d'obtention, le microbiote intestinal de termite apparait donc comme une source prometteuse d'inoculum pour la conversion de lignocellulose en acides gras volatils.

L'identification des bactéries composant ce microbiote a permis de confirmer la dominance *Spirochaetes, Fibrobacteres,* et *TG3* pour les trois espèces testées ayant un régime alimentaire à base de bois, la quatrième, humivore, présentant un profil plus équilibré, partagé entre *Bacteroidetes, Firmicutes, Proteobacteria, Spirochaetes* et *Actinobacteria*. Cette dernière est celle dont le microbiote s'est le moins bien acclimaté aux conditions du réacteur, présentant un faible taux de dégradation. Pour les autres, présentant des meilleures capacités de dégradation, montrent cependant une évolution drastique des communautés microbiennes qui les composent. En quinze jours de culture, les trois phyla majoritaires dans l'intestin disparaissent quasiment complètement ; ils sont remplacés par des communautés dominées par des *Firmicutes et Proteobacteria*, parfois associées à des *Bacteroidetes*. Cependant, si les communautés initiales n'ont pas été maintenues, les communautés qui les ont remplacées présentent tout de même de très bonnes capacités de dégradation.

Au vu des variations de diversité observées, il est donc nécessaire de passer par une étape de stabilisation par enrichissement pour poursuivre l'utilisation de microbiote de termite en fermenteur. Celle-ci pourrait de plus aboutir à une augmentation des capacités de dégradation, comme observée avec le rumen. Elle permettra aussi d'étudier l'impact de l'inoculum de départ sur la dégradation d'un même substrat, ainsi que sur la communauté finale obtenue après enrichissement.

CHAPITRE VII

ANALYSE DE L'EVOLUTION DE LA DIVERSITE AU COURS DE L'ENRICHISSEMENT A PARTIR DE MICROBIOMME INTESTINAL DE TERMITES

CHAPITRE VII : ANALYSE DE L'EVOLUTION DE LA DIVERSITE AU COURS DE L'ENRICHISSEMENT A PARTIR DE MICROBIOME INTESTINAL DE TERMITES

IV. Introduction

Dans le chapitre précédent, la mise en culture du microbiote intestinal de quatre espèces de termites a permis de démontrer que ceux-ci, et plus spécialement le microbiote de *Nasutitermes ephratae*, pouvait être une source d'inoculum intéressante pour la conversion de lignocellulose en carboxylates. Cependant, si des performances de dégradation élevées ont été observées, les communautés bactériennes associées ont présenté des changements extrêmement forts et une variabilité entre réplicats forte également. Ces communautés doivent donc être stabilisées.

La stabilisation des communautés peut être obtenue par enrichissement en réacteurs successifs, comme cela a été réalisé avec un inoculum rumen dans le chapitre IV. Dans le cas du rumen, cette stabilisation s'était de plus accompagnée d'une augmentation des capacités de dégradation de la lignocellulose. Celles-ci étant déjà très élevées au premier cycle de dégradation avec un inoculum termite, leur maintien serait déjà une réussite. Par ailleurs, l'enrichissement d'une communauté à partir de microbiote de termite peut permettre d'étudier la part relative des effets « inoculum » et « substrat » sur la composition finale des communautés obtenues. Plus celles-ci seront proches de celles obtenues à partir de rumen bovin, plus le substrat a un effet fort sur la sélection et le façonnage de la communauté capable de le dégrader. Dans des études antérieures, l'effet substrat a déjà été décrit comme plus fort que l'effet inoculum, mais les inoculum utilisés provenaient de compost et étaient relativement proches entre eux (Simmons et al., 2014). Dans le cas des différences très grandes entre microbiote ruminal bovin et intestinal termite, l'effet inoculum peut être plus fort que l'effet substrat.

Dans ce chapitre, les communautés précédemment obtenues après un cycle de dégradation à partir d'inoculum intestinal du termite *Nasutitermes ephratae* ont suivi un même processus d'enrichissement en utilisant en parallèle un substrat stérile ou non stérile, afin de limiter l'apport de bactéries endogènes du substrat ou exogènes, provenant de l'environnement. Les communautés obtenues ont ensuite été caractérisée en cinétique au cours d'un cycle de dégradation de la paille de blé. Ceci a permis d'identifier les taxons majoritaires dont l'activité apparait liée à la dégradation. Finalement, cette dynamique a été comparée à celle observée précédemment dans la communauté issue du rumen.

V. Dynamique des communautés d'un consortium bactérien dérivé du microbiote intestinal de termite.

Ce chapitre est présenté sous la forme d'un article en préparation :

Community dynamics of a lignocellulolytic bacterial consortium derived from termite-gut microbiome.

Lucas Auer[1,2,3], Adèle Lazuka[1,2,3], Michael O'Donohue[1,2,3], Guillermina Hernandez-Raquet[1,2,3*]

[1)] Université de Toulouse, INSA, UPS, LISBP, 135 Avenue de Rangueil, F-31077 Toulouse, France
[2)] INRA, UMR792 Ingénierie des Systèmes Biologiques et des Procédés, F-31400 Toulouse, France
[3)] CNRS, UMR5504, F-31400 Toulouse, France

V.1. Introduction

Lignocellulose (LC) is the main component of plant cell wall, and is one of the most abundant polysaccharides on Earth. In the current context of sustainable development, the biorefinery of non-food lignocellulosic agricultural residues into chemicals is a promising way to reduce fossil carbon dependency. However, LC is a substrate difficult to degrade. It is made of three main components, cellulose, hemicelluloses, and lignin which are associated to each other in a complex structure that confers to LC resistance to biotic and abiotic attack (Menon and Rao, 2012). In the current context of LC valorization via bioethanol production, LC bioconversion involves classically the use of physico-chemical pretreatments, followed of enzymatic hydrolysis to produce monomeric sugars (sugar platform) which are further fermented by selected yeast strains under sterile conditions. However, the difficulty to implement sustainable pretreatment methods and the limited efficiency of enzymatic cocktails for biomass deconstruction are the main bottlenecks to develop such LC biorefinery. A second approach for lignocellulosic biomass valorization already in use is the methane production. It relies on microbial mixed community's metabolism to transforms LC under anaerobic conditions; such process is also called the biogas platform. Here, part of the substrate is converted into microbial biomass, but most of it is fermented into carboxylates, mainly short chain fatty acids (VFA), which are further converted into biogas. The metabolism of microbial communities could also be oriented to accumulate VFAs; this bioprocess is named the carboxylate platform. Carboxylate production is of particular interest because they can be used as synthons for further biological transformation to produce added value products such as e.g polyhydroxy-alkanoate bioplastics. Thus, the selection and

stabilization of microbial communities that efficiently convert LC into carboxylates are the key parameters to develop an efficient LC bioconversion processes for both methane and carboxylate production (Agler et al., 2011).

In this context, several studies reported successful LC degradation using microbial communities (Feng et al., 2011; Gao et al., 2013; Eichorst et al., 2014). However, the reported LC transformation rates are generally low. Conversion rates could be improved by finding more efficient microbial biocatalysts. Although LC biotransformation is currently a challenge for biotechnology, it is also a widespread process naturally occurring in most of the terrestrial and aquatic ecosystems. The recycling of carbon stored in plant biomass is realized in soils and aquatic sediments thanks to the action of aerobic fungi and bacteria which produce extracellular hydrolytic enzymes; such enzymes are often used in the sugar platform (Klein-Marcuschamer et al., 2012). In anaerobic or anoxic ecosystems, such as wetlands or peatlands, LC degradation can results in methane production by Archaea. In animals which consume plants as main source of food, lignocellulose transformation is mainly carried out by the microbiota present in the digestive tract. Indeed, most animal are not enzymatically equipped to degrade lignocellulose; its capacity to transform biomass relies on their microbial symbionts for its digestion into VFAs which are then assimilated by the host. Herbivorous animals, such as mammalian ruminants, display very high degradation rates. Similarly, insects, and particularly termites, are able to degrade lignin-rich lignocellulosic substrates such as wood. Such digestive systems also produce methane, but this production is generally based on H_2 and CO_2 consumption which does not compete with VFA production. However, despite the large diversity of lignocellulolytic ecosystems and its underlying diversity of LC deconstruction processes, most of them have not been explored for their potential to transform lignocellulose in controlled conditions, particularly those were anaerobic conditions prevails.

Termites are the best LC degraders, able to remove up to 99% cellulose and 87% hemicellulose of wood (Brune, 2014). In LC digestion by termites, the host plays a role via mechanical (chewing) and chemical actions (a basic pH prevailing in the digestive tract), but most of the biomass hydrolysis relies on its gut microbiome (Brune, 2014). The termite digestive system and the diversity of its microflora have been well characterized: LC degradation in lower termites relies on symbiotic interactions between the host, eukaryotic flagellates and bacteria whereas in higher termites of the *Macrotermitinae* family LC transformation occurs thanks to an external symbiosis with fungi, so these last degrade LC

prior its ingestion by termites. In other phylogenetic groups of higher termites, LC digestion relies on a gut microflora exclusively composed of prokaryotes mainly belonging to *Spirochaetes*, *Fibrobacteres* and new uncharacterized phyla (Hongoh et al., 2006). The diversity of such gut microbiome strongly differs from that observed in herbivorous mammalian or soil communities, mainly constituted by *Firmicutes* and *Bacteroidetes*. In a previous study, we demonstrated that the gut microbiome of four different higher termite species were able to transform lignocellulosic biomass in controlled bioreactors (Auer et al., 2016, in preparation). The gut-microbiome of *Nasutitermes ephratae* displayed the highest LC degradation and VFA production rates. Here, our aim was to assess the potential of *N. ephratae* derived microflora to be implemented, enriched and maintained in controlled bioreactors to produce carboxylates from lignocellulosic biomass. The microbial diversity during the enrichment process using wheat straw as sole carbon source was characterized by 16S rRNA gene sequencing. In order to assess the robustness of the community, we compared the enrichment processes using a sterilized and non–sterilized LC substrate. Finally, the community dynamics during a complete lignocellulose degradation cycle was compared to that observed with a cow rumen-derived inoculum, stabilized under similar conditions. Despite a strong shift on the community diversity was observed during the termite gut enrichment processes, our results provide evidence that a termite-derived consortium maintained its lignocellulose transformation capacity in the conditions imposed in bioreactors.

V.2. Methods

V.2.1 Enrichment of termite-gut community in lignocellulose transformation reactors

The enrichment of the termite-gut community and its kinetic characterization were performed in anaerobic batch reactors (2L BIOSTAT® A+, Sartorius, Germany) using a mineral medium described by Lazuka et al., (2015) and wheat straw as the sole carbon source (Koreli variety, 20g.L^{-1}, 2 mm size). For the enrichment of the termite-gut community a sequential batch reactor approach was used. Due to the low amount of inoculum available (1,000 guts), the first enrichment reactor was realized in a small volume reactor (400mL, Applikon Minibio 500) and inoculated directly with the dissected termite guts, as previously described (Auer et al., 2016, in preparation). At the end of this first cycle of cultivation (Cycle 1), the total volume of the reactor was used to inoculate the next one. Thereafter, each

successive reactor was inoculated (10% v/v) with the end of the previous culture. Two parallel enrichment reactors were performed using sterilized (dry autoclaving, 121°C, 1.3 bar) or non-sterilized wheat straw. All reactors were conducted in biological duplicates at 35°C, controlled pH (6.15), under aseptic conditions.

During the enrichment process and for the kinetics characterization, lignocellulose degradation was determined considering the total solids (TS) and volatile solid (VS) measured in 10 mL samples withdrawn along the incubation time. These samples were first centrifuged (7197xg, 10 min), rinsed twice with distilled water and dried 24 h at 105°C (TS). The mineral fraction (MF) was estimated by mineralization of the samples at 500°C for 2 h; VS were estimated from the difference between TS and MF. Lignocellulose degradation was equivalent to VS degradation and expressed as percentage (w/w).

V.2.2 Diversity analyses

Bacterial diversity was assessed at each initial and final point during the enrichment processes and daily for the kinetic characterization reactors. For DNA and RNA extraction, 1.5mL samples were taken and immediately centrifuged (13,000 x g, 5min and 4 °C), the supernatant was removed and the pellet was snap frozen in liquid nitrogen. Samples were stored at -80°C until nucleic acid extraction. Total DNA and RNA were extracted using a PowerMicrobiome RNA Isolation kit (MoBio Laboratories Inc. Carlsbad) following the manufacturer's instructions but omitting the final DNAse steps. Cell lysis was carried out with a Fast Prep (MP Biomedicals) (2 x 30s at $4ms^{-1}$). Subsequently, DNA and RNA were separated and purified using an AllPrep DNA/RNA Mini Kit (Qiagen) according to the manufacturer's instructions. DNA and RNA integrity and purity were checked by agarose gel (1%) electrophoresis. Concentrations were measured by NanoDrop 1000 spectrophotometer (Thermo Scientific), measuring absorbance at 260 and 280 nm. Residual DNA content in RNA samples was removed using 1µg RNA and a DNAse (TURBO DNA-free™ Ambion, Life Technologies) according to the manufacturer's instructions. RNA was then retro-transcripted into cDNA using M-MLV Reverse Transcriptase (Promega) and random hexamers (Roche) following the manufacturer's instructions. Illumina sequencing of the V3-V4 region of 16S rRNA gene was performed after PCR amplification using the bacterial primers 343F and 784R, modified to add sequencing adaptors during a second PCR (343F= 5'-CTT TCC CTA CAC GAC GCT CTT CCG ATC TAC GGR AGG CAG CAG-3' and

784R= 5'-GGA GTT CAG ACG TGT GCT CTT CCG ATC TTA CCA GGG TAT CTA ATC CT-3'). Library was prepared as previously detailed (Auer et al., 2016, in preparation), and loaded on a MiSeq Illumina cartridge, using reagent kit v3 (paired 300bp reads). Sequencing was performed at the GenoToul Genomics and Transcriptomics facility (GeT, Auzeville, France) using a MiSeq® Illumina®.

V.2.3 Data processing

Illumina sequencing data were demultiplexed and pair-end reads were joined by the GeT platform, using Flash v1.2.6 (Magoč and Salzberg, 2011), 110bp minimum overlap and a 0.1 maximum mismatches ratio. Fastq files were transformed into a unique fasta file and treated with Mothur v1.33.1 (Schloss et al., 2009) following the SRF1 procedure described previously (Auer et al., submitted). A fusion of LTP SSU database (version 115, Yarza et al., 2008) and DictDB (Mikaelyan et al., 2015), a database dedicated to insect-associated bacteria was used to improve taxonomic assignation of sequences from termite gut-derived communities. Phylogenetic trees were constructed using ClustalO (Sievers et al., 2014) and raxmlHPC (Stamatakis, 2014). OTU tables, taxonomic files and phylogenetic trees were imported into Phyloseq package v1.14.0 (McMurdie and Holmes, 2013) using R v3.2.3, following the manual import procedure. Shannon and Simpson indexes, Bray-Curtis and weighted-Unifrac distances between samples and Principle Coordinate Analysis (PCoA) ordinations were all performed using Phyloseq functions. Weighted-Unifrac distances were clustered using *hclust* function and ward.D2 method of the R v3.2.3 *stats* package. Partial Least Square (PLS) analyses were performed using mixOmics v5.2.0 package (Cao et al., 2009). ANOVA analyses were performed using R functions *anova* and *lm* and PERMANOVA analyses were performed using *adonis* function of the *vegan* package v2.3.5.

V.3. Results

V.3.1 Termite-gut microbiome enrichment

The percentages of LC degradation determined at each cycle of the enrichment of *Nasutitermes ephratae* gut-microbiome, using sterilized and non-sterilized wheat straw, are summarized in Table 1. When non-sterilized straw was used, the percentage of LC degradation continuously decreased from 40% at Cycle 2 to 26% at Cycle 6. In contrast, when cultures were realized with sterilized straw, degradation stabilized around 38% from Cycle 4 to the end of the enrichment process (Cycle 6).

Table 1: Average lignocellulose degradation measured along the enrichment process on sterile or non-sterile straw.

	Cycle	Degradation (w/w)
Sterile Straw	C1	45±4%
	C2	40±4%
	C3	32±2%
	C4	33±4%
	C5	27±3%
	C6	26±4%
Non-Sterile Straw	C2	48±5%
	C3	44±2%
	C4	38±0,4%
	C5	38±1%
	C6	37±2%

Bacterial diversity analysis along the enrichment using both sterilized and non-sterilized straw showed that Simpson and Shannon diversity index of diversity were highly close between replicates (Table 2). For both sterilized and non-sterilized substrates a strong decrease on diversity was observed at Cycle 2. Thereafter, diversity index stabilized at slightly higher values. There were no significant differences between the diversity estimates for enrichments performed with sterilized or non-sterilized straw.

Table 2: Change on diversity index along the enrichment process with sterile or non-sterile wheat straw. For comparison, the diversity along the enrichment of a rumen-derived inoculum is also showed (Lazuka et al., 2015).

			Diversity index		
Enrichment	Substrate	Cycle	Richness	Shannon	Simpson
Termite Gut	Sterile	Inoculum	171±1	3,3±0,1	12±1,4
		C1	63±12	2,2±0,7	6±0,9
		C2	118	1,1	2,0
		C3	119±8	2±0,03	4±0,1
		C4	132±5	2±0,1	5±0,03
		C5	109±11	2±0,1	5±0,2
		C6	121±6	1,9±0,1	4±0,3
	Non-sterile	C2	164	1,6	2,0
		C3	123±1	1,9±0,1	3±0,1
		C4	122±1	2±0,07	4±0,2
		C5	119±7	2±0,07	4±0,4
		C6	127±10	1,9±0,03	4±0,3
Rumen	Non-sterile	Inoculum	413±15	4,6±0,2	35±9
		C1	285±8	3,2±0,2	14±2
		C2	245	2,9	10,0
		C3	231	3,1	11,8
		C4	218	3,1	11,2
		C5	216	2,7	6,2
		C6	207	2,7	6,0
		C7	214	2,8	7,1
		C8	216	2,9	8,2
		C9	171	2,7	7,0
		C10	176	3,0	8,8

Microbial community profiles showed that the *Nasutitermes gut* inocula, initially dominated by *Spirochaetes* and *Fibrobacteres*, shifted after the first cycle of enrichment on wheat straw, being largely dominated by *Firmicutes* (Figure 1, C1). From Cycle 2, the abundance of members related to *Bacteroides* and *Prevotella* genus (*Bacteroidetes* phyla) strongly increased at the expenses of *Firmicutes*, with both sterilized and non-sterilized wheat straw. The following enrichment cycles presented a similar and stable microbial community composition, with small variations of *Bacteroidetes* and *Firmicutes* content. Nevertheless, in the final enriched communities (Cycle 4 to Cycle 6), some genera were specific to a given type of straw. For instance, *Butyrivibrio* reached 10% when non-sterile straw was used whereas its abundance was much lower when sterilized straw was utilized. At the opposite,

unclassified *Lachnospiraceae* and *Rikenellaceae RC9* genus were more abundant when sterile straw was used compared to non-sterile straw.

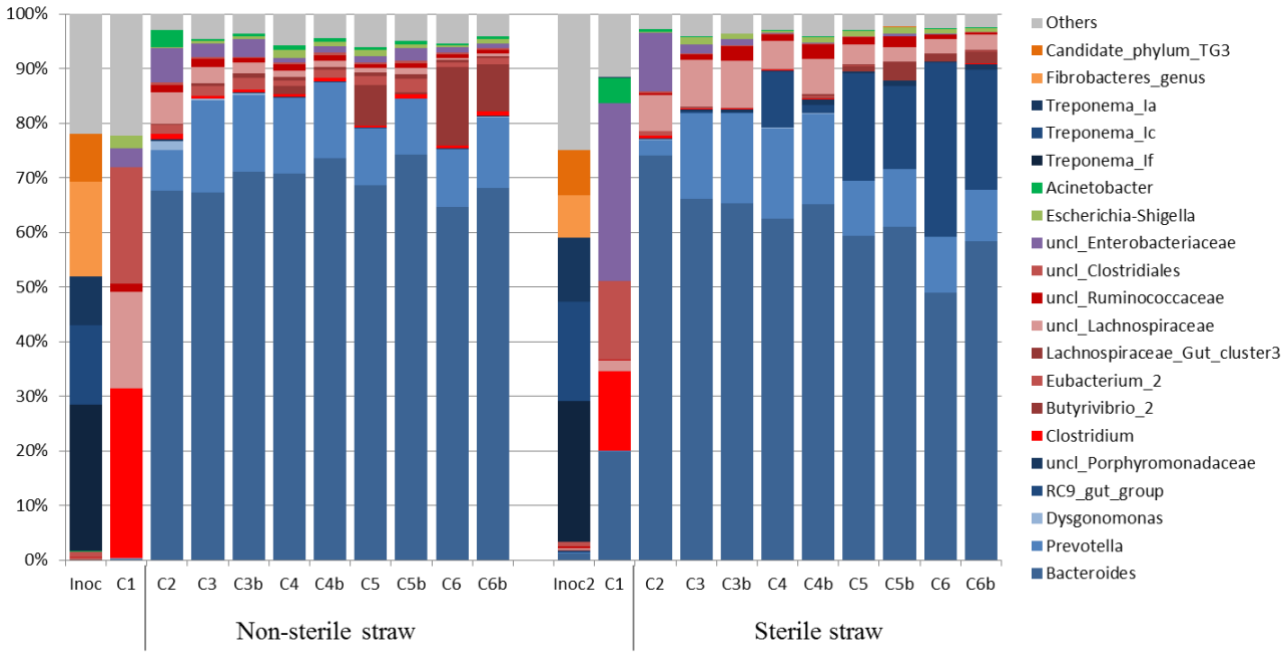


Figure 1: Microbial community composition during the enrichment of *N. ephratae* gut microbiome (Inoc) using non-sterile (left) and sterile straw (right). Cycles 3 to 6 were performed in duplicates (C_n and C_nb). Genii are colored according to their phylum: *Bacteroidetes* in blue, *Firmicutes* in red, *Proteobacteria* in green, *Spirochaetes* in purple, and *Fibrobacteres* and *TG3* in orange.

The diversity data obtained from the termite-gut enrichments on sterilized and non-sterilized wheat straw were compared with previous data obtained from a cow-rumen enrichment conducted in the same conditions and showing similar lignocellulose degradation capacities (Lazuka et al., 2015). Principle Coordinate Analysis based on Bray-Curtis distances (Supplementary data, Figure 1A) and clustering of weighted-Unifrac distances (Supplementary data, Figure 1B) confirmed the presence of a diversity switch between the initial inocula and their respective enrichment. The final communities enriched from the termite-gut realized with sterile and non-sterile straw clustered together, whereas their respective inoculum source clearly constituted the out-group. During the enrichment processes, it is possible to remark that termite-derived communities were significantly distant to the cow-rumen enrichment. However, the final termite-gut communities obtained after the enrichment process were less distant to the final cow-rumen derived communities than to the initial termite gut inocula. Microbial communities derived from a termite-gut inoculum were thus more similar to rumen-derivate communities than to the original termite-gut inoculum.

V.3.2 Diversity dynamics throughout lignocellulose degradation by TWS

The final termite-gut derived consortium obtained after enrichment on sterile straw (named TWS), displaying high lignocellulose degradation levels, was selected for further kinetics characterization. A lignocellulose degradation cycle was conducted in duplicate, named ST1 and ST2, using the same culture conditions previously described and realizing a frequent sampling along the incubation. Such detailed kinetic analysis enable to better characterize the substrate degradation rate, characterize the community composition based on the 16S rDNA gene content, and identify the main metabolically active taxa based on the 16S rRNA expression level (RNA) (Blazewicz et al., 2013). In this way, it is possible correlate the changes on the metabolically activity bacteria (at the genus level) with the LC degradation dynamics.

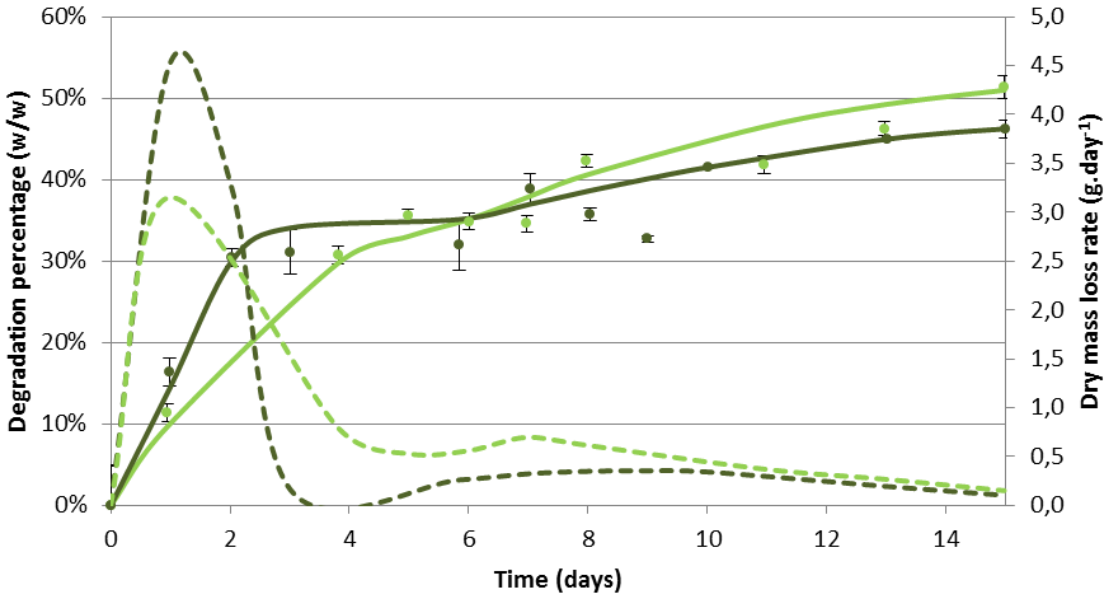


Figure 2: Lignocellulose degradation and its corresponding degradation rate during cultivation of TWS in wheat straw. Biological duplicates ST1 and ST2 are colored in light and dark green, respectively. Full lines represent the percentage of LC degradation, with dots corresponding to experimental points with their associated standard deviation. Doted lines represent the dry mass loss rate.

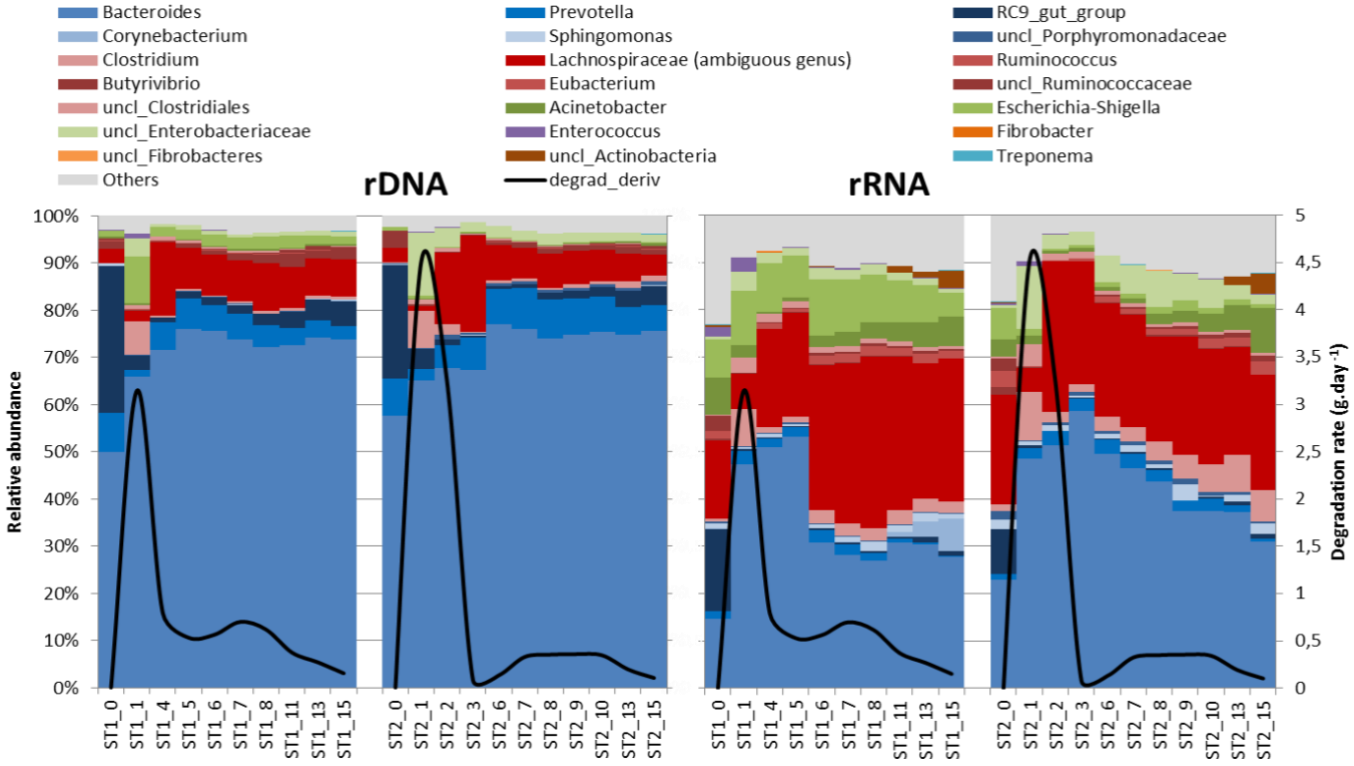


Figure 3: Changes in the microbial community composition (16S rDNA, left) and metabolically active populations (16S rRNA, right) during lignocellulose degradation. Numbers in X-axis corresponds to incubation days. Colors correspond to bacterial classes ordered by phyla : *Bacteroidia* (blue), *Clostridia* (red), *Proteobacteria* (green), *Bacilli* (purple), *Fibrobacteres* (orange), *Actinobacteria* (brown) and *Spirochaetes* (cyan). Lignocellulose degradation rate is indicated with a black curve.

LC degradation during the fermentation period reached 46 to 52% after 15 days of incubation (Figure 2). The LC degradation rate was variable during time, displaying an initial high LC degradation phase, between day 1 and 4; thereafter the LC degradation rate decreased. During this phase, in both biological replicates, *Bacteroides* related members were the most abundant, reaching 80% of total 16S rDNA sequences (Figure 3). *Rikenellaceae RC9* relatives, highly abundant in the inocula, were replaced by *Proteobacteria* members since day one, and then by *Lachnospiraceae*. The metabolically active microbial groups, as determined by 16rRNA expression profiles, showed a similar behavior displaying the same dominant groups but with less contrasted abundances than those observed for 16S rDNA profiles. The expression level of *Bacteroides* members was less marked compared to its relative abundance determined from 16S rDNA data, reaching a maximum level of about 60% during the first five days of incubation; thereafter the expression of this group decreased, being replaced by *Lachnospiraceae*, *Proteobacteria* and several minor bacterial groups (representing

individually less than 1% of the community and grouped as "Others" in Figure 3). Compared to the 16S rDNA abundance profile, the expression level of *Lachnospiraceae*, *Proteobacteria* and *Others* was higher. *Bacteroides* abundance was high during the strong LC degradation phase considering both 16S rDNA (up to 80%) and rRNA (up to 70%). After this period of strong lignocellulolytic activity, the community composition was relatively stable. Principle Coordinate Analysis (PCoA) based on weighted-Unifrac distances confirmed that the community composition markedly changed during the first days of culture (Supplementary data 2) and stabilized after few days of incubation. This behavior differs slightly when considering 16S rDNA or rRNA data: indeed, while the community composition based on 16S rDNA showed minor changes after 5 days of incubation, the diversity of metabolically active bacteria (based on 16S rRNA) reached a stable state slightly later.

V.3.3 Comparing the diversity dynamics of termite-derived and rumen-derived consortia along lignocellulose transformation.

The dynamics of TWS microbial diversity during LC degradation was compared to that observed in experiments conducted under similar conditions with a rumen-derived inoculum (named RWS), detailed in previous studies of our group (Lazuka et al., 2015, Auer et al., in preparation). The dynamics of the metabolically active community (rRNA data) of RWS during LC degradation is shown in Supplementary data 3. It is possible to remark that the communities are quite similar, but also present strong differences. Indeed, the active genii in TWS (*Bacteroides* and *Lachnospiraceae*) were also present and active in RWS, even if they were less dominant: RWS presented a much higher diversity of genii, and their activity was quite well balanced. In order to identify the main groups involved on LC degradation in TWS and RWS, the dynamics of 16S rRNA sequencing data corresponding to the initial, maximal (peak) and plateau of LC degradation by both inocula were compared using Weighted-Unifrac distances (Figure 4).

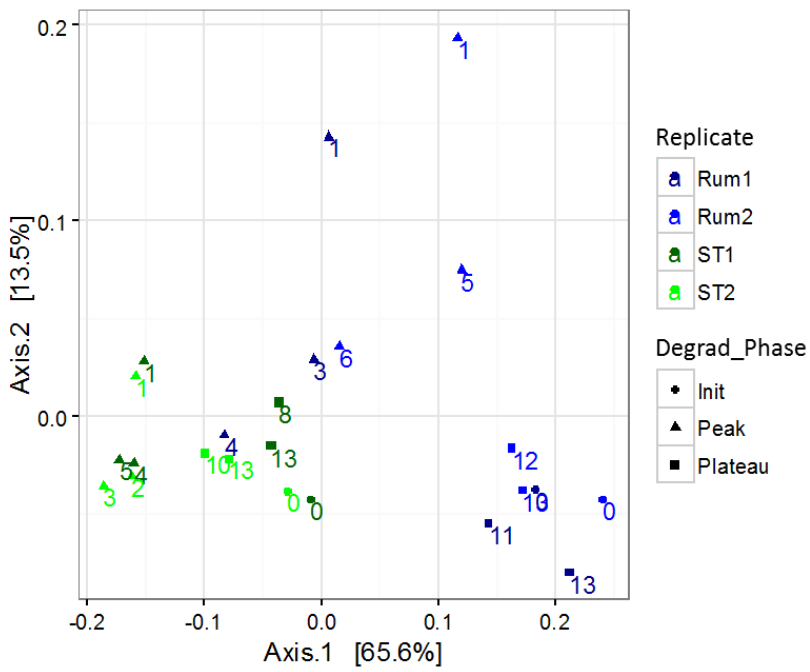


Figure 4: PCoA analyses of TWS and RWS dynamics along the LC degradation cycle (weighted-Unifrac distances). Rum1 and Rum2 correspond to the RWS replicates (in blue), ST1 and ST2 indicate the TWS replicates (in green). The numbers indicate the incubation time (days) and the symbols indicate the three lignocellulose degradation phases.

RWS and TWS were well separated and presented the same type of dynamics: for each inoculum, the communities corresponding to the initial and plateau phases of LC degradation were regrouped together whereas the community corresponding to the peak of lignocellulolytic activity were more dissimilar, indicating that the active community associated with the LC degradation peak differed from initial and final states. A PERMANOVA analysis (Table 3) showed that the "type of inoculum" (42%), the "degradation phases" (13%) and their "interaction" (5%) were the main factors explaining variance between samples while the "replicates" had no significant effect.

Table 3: PERMANOVA analysis based on weighted-Unifrac distances, using "inoculum", "degradation phase", "replicate", and "interactions" between them as factors. The degrees of freedom (Df), sequential sums of squares (SumsOfSqs), mean squares (MeanSqs), *F* statistics, partial *R-squared* and *P* values based on 9999 permutations are indicated.

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)	
Inoculum	1	0.28	0.28	**41.7**	0.43	0.0001	***
Degrad_Phase	2	0.17	0.09	**12.8**	0.26	0.0001	***
Replicate	2	0.02	0.01	1.8	0.04	0.1303	
Inoculum:Degrad_Phase	2	0.06	0.03	**4.5**	0.09	0.0068	**
Degrad_Phase:Replicate	4	0.03	0.01	1.2	0.05	0.3219	
Residuals	12	0.08	0.01		0.12		
Total	23	0.66			1.0		

V.3.4 Identification of OTUs common and specific to TWS and RWS

At the OTU level, a sparse Partial Least Square Discriminant Analysis (sPLS-DA) approach (Figure 5) using the metabolically active OTUs (16S rRNA) data, and "inoculum" as the discriminant factor clearly separated RWS and TWS communities according to the X-axis. Among the 15 selected discriminant OTUs, OTU_1, OTU_3 and OTU_{26} appeared to be characteristic of TWS whereas the other discriminant OTUs were mainly expressed in RWS. ANOVA analyses (Table 4) of the selected OTUs confirmed that OTU_1, OTU_3 and OTU_{26}, related to *Bacteroides*, *Lachnospiraceae* and *Sphingomonas*, respectively, were significantly more expressed in TWS compared to RWS consortium. At the opposite, OTU_{21}, OTU_{25}, OTU_{31}, OTU_{32}, OTU_{36} and OTU_{49}, related to *Dysgonomonas*, *Fibrobacter*, *Actinobacteria*, *Lactobacillales*, *Lachnospiraceae* and *Rumonicoccaceae*, respectively, were significantly more active in RWS compared to TWS. The other OTUs identified with the sPLS-DA approach were not fully statistically supported by ANOVA analysis. Component 2 separated vertically the communities corresponding to RWS plateau from the other RWS samples, capturing part of the temporal variation observed between RWS samples; however, such temporal trend was not observed for TWS.

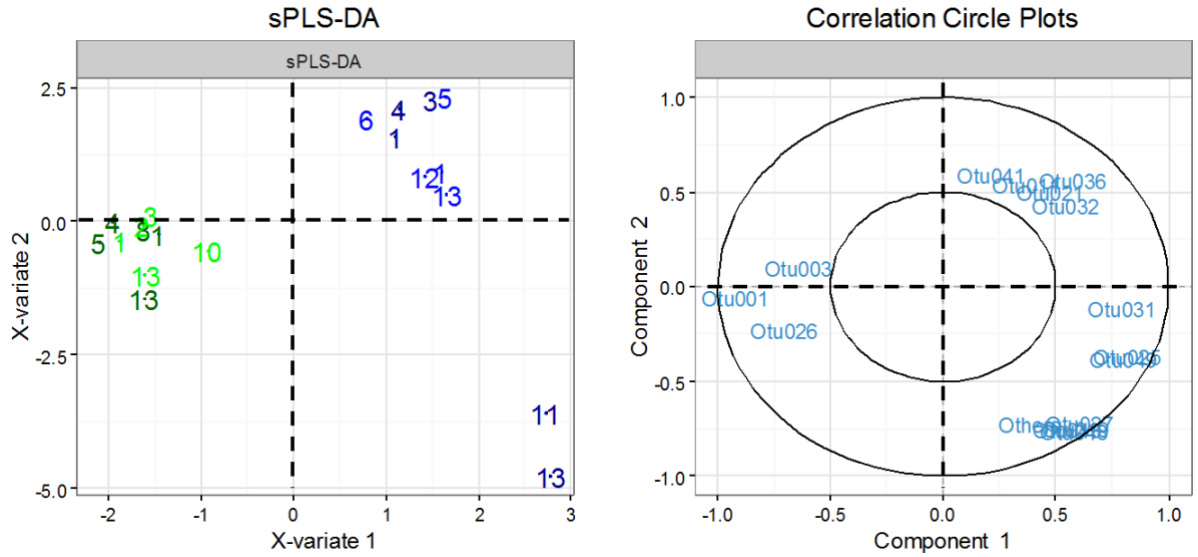


Figure 5: sPLS-DA analysis based on 16S rRNA data of representative points (1 for initial, 3 for degradation peak and 2 for plateau, for each replicate) and "inoculum" as discriminant factor. On the left, sPLS-DA plot of RWS (blue) and TWS (green) active communities, labelled according to the incubation time (in days). The corresponding correlation circle is shown on the right.

Table 4: Average rRNA level of OTUs identified by sPLS-DA in RWS and TWS. Analysis is based on representative samples (6 samples per replicate: 1 initial, 3 during degradation peak and 2 during the final plateau of LC degradation). OUTs taxonomy at the Class level and at the most precise known level of taxonomy are indicated. ANOVA values and significance of differences observed between RWS and TWS are detailed on the right.

			Mean rRNA relative abondance		ANOVA	
OTU	Class	Taxonomy	RWS	TWS	P-value	
1	*Bacteroidia*	*Bacteroides*	9,86%	37,58%	1,38E-07	***
3	*Clostridia*	*uncl Lachnospiraceae*	6,95%	22,19%	0,002714	**
11	*Bacteroidia*	*Prevotella*	0,83%	0,12%	0,02345	*
19	*Actinobacteria*	*Corynebacterium*	2,53%	0,38%	0,08394	.
21	*Bacteroidia*	*Dysgonomonas*	0,95%	0,08%	0,004839	**
25	*Clostridia*	*uncl Lachnospiraceae*	0,36%	0,00%	0,0004853	***
26	*α-proteobacteria*	*Sphingomonas*	0,00%	0,89%	0,0003055	***
27	*Bacteroidia*	*uncl Porphyromonadaceae*	0,50%	0,02%	0,06717	.
31	*Actinobacteria*	*Actinobacteria*	0,56%	0,00%	0,0001878	***
32	*Bacilli*	*Lactobacillales*	1,10%	0,00%	0,002876	**
36	*Clostridia*	*uncl Lachnospiraceae*	0,28%	0,00%	0,001323	**
41	*Clostridia*	*uncl Clostridia*	0,45%	0,23%	0,1271	
43	*Spirochaetes*	*uncl Spirochaetales*	0,93%	0,03%	0,1147	
48	*Bacteroidia*	*Rikenellaceae RC9 gut group*	0,03%	0,00%	0,1054	
49	*Clostridia*	*uncl Ruminococcaceae*	0,36%	0,00%	0,001072	**
		Others	9,71%	7,56%	0,3831	

In order to identify OTUs specific to a given phase of LC degradation and inoculum source, a sPLS-DA analysis was performed using "inoculum" and "degradation phase" as discriminant factors (Figure 6). Such analysis enabled to clearly separate the samples corresponding to the peak and plateau of LC degradation of RWS, while TWS samples were less well discriminated by the "degradation phases". It was possible to confirm that high expression levels of OTU_1 and OTU_{26} were characteristics of TWS while twelve OTUs were associated to RWS, OTU_1, OTU_7, and OTU_{21} were related to the LC degradation peak and OTU_5, OTU_{16}, OTU_{18}, OTU_{19} to the plateau of LC transformation. ANOVA analysis of the sPLS-DA discriminant OTUs and the comparison of their mean 16S rRNA relative abundance (Table 5) allowed identifying some important OTUs: OTU_1, related to *Bacteroides*, was more active in TWS than in RWS, and displayed strong differences between the degradation peak and plateau samples (43% of the total 16S rRNA content versus 29%). OTU_{26}, belonging to *Sphingomonas*, was specific of TWS, with a two-fold expression increase between the LC degradation peak and the plateau, but remained a relatively minor OTU (representing 0.6% to

1.3% of total 16S rRNA content). The other major OTUs of TWS were not well separated due to the high variability of RWS. Category "Others", which contained all minor OTUs, displayed an increased relative activity during the plateau of LC degradation, cumulating a mean of 11% of abundance whereas its abundance never exceeded 5% during the LC degradation peak. *Fibrobacteres* OTUs (OTU_{16}, OTU_{22}) were identified to be especially active in RWS during the plateau of LC degradation; this group was not detected in TWS.

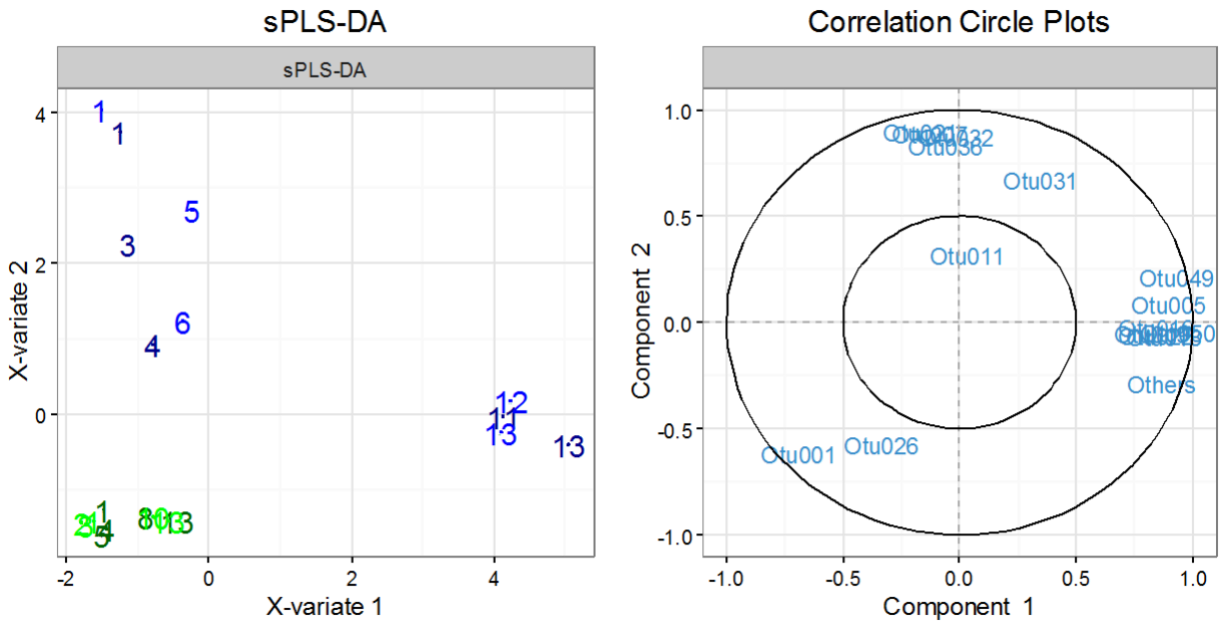


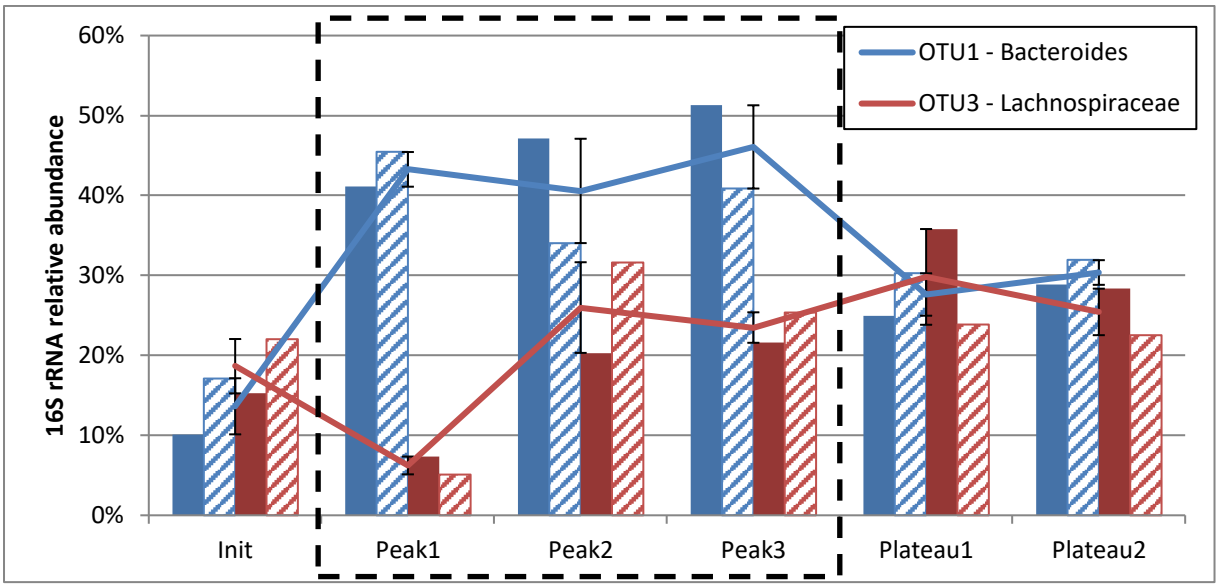
Figure 6: sPLS-DA analysis based on 16S rRNA data of representative points (1 point for initial, 3 for degradation peak and 2 for plateau, for each replicate) using "degradation phases" and "inoculum" as discriminant factors. On the left, sPLS-DA plot of RWS (blue) and TWS (green) active communities, labelled according to the incubation time (in days). The corresponding correlation circle is shown on the right.

Table 5: Mean rRNA levels of OTUs identified by sPLS-DA during the peak and plateau phases of LC degradation by RWS and TWS. Analysis is based on representative samples (5 samples per replicate: 3 during degradation peak and 2 during final plateau). OTUs taxonomy at the Class level and at the most precise level of taxonomy known are indicated. ANOVA values and significance of differences observed between RWS and TWS are detailed on the right.

			Mean rRNA relative abundance				ANOVA	
			RWS		TWS			
OTU	Class	Taxonomy	Peak	Plateau	Peak	Plateau	P-value	
1	*Bacteroidia*	*Bacteroides*	13,29%	4,72%	43,30%	28,99%	1,49E-09	***
5	*ɣ-proteobacteria*	*Acinetobacter*	3,68%	9,34%	0,76%	3,76%	2,39E-06	***
7	*Bacilli*	*uncl Lactobacillales*	16,44%	3,66%	0,74%	0,12%	0,006715	**
11	*Bacteroidia*	*Prevotella*	1,11%	0,41%	0,17%	0,06%	0,05178	.
16	*Fibrobacteres*	*uncl Fibrobacteres*	0,04%	7,41%	0,03%	0,00%	1,97E-05	***
18	*Clostridia*	*Eubacterium*	1,01%	6,40%	0,42%	1,93%	9,99E-06	***
19	*Actinobacteria*	*Corynebacterium*	0,30%	5,89%	0,05%	0,86%	0,0003862	***
21	*Bacteroidia*	*Dysgonomonas*	1,40%	0,28%	0,10%	0,06%	0,0005427	***
22	*Fibrobacteres*	*uncl Fibrobacteres*	0,01%	6,96%	0,01%	0,00%	0,0002852	***
26	*α-proteobacteria*	*Sphingomonas*	0,00%	0,00%	0,61%	1,33%	0,000253	***
31	*Actinobacteria*	*uncl Actinobacteria*	0,57%	0,54%	0,00%	0,00%	0,000253	***
32	*Bacilli*	*uncl Lactobacillales*	1,49%	0,53%	0,00%	0,00%	0,004864	**
36	*Clostridia*	*uncl Lachnospiraceae*	0,41%	0,08%	0,00%	0,00%	3,68E-05	***
49	*Clostridia*	*uncl Ruminococcaceae*	0,16%	0,67%	0,00%	0,00%	9,57E-08	***
50	*Clostridia*	*uncl Ruminococcaceae*	0,01%	1,78%	0,00%	0,00%	6,12E-07	***
		Others	4,83%	17,05%	4,97%	11,44%	1,18E-07	***

In order to confirm the impact of LC degradation phases on the expression of different microbial groups, sPLS-DA analysis was performed exclusively on TWS data (Supplementary data 4). It confirmed that OTU_1 was particularly active during the LC degradation peak. OTU_3 appeared to be mainly active during the plateau phase, with an average of 28% abundance of the total 16S rRNA content versus 19% during the peak of LC transformation; however, such differences were not statistically supported by supervised classification methods. Indeed, OTU_3 displayed a high variability during the peak of LC degradation (Figure 7), due to a delay of its activity increase. It in fact presented a contrasted activity dynamic, but its peak was not synchronized with the OTU_1 and degradation peak. It also presented a sustained activity from its peak to the plateau, while the activity of other major OTUs decreased.

Figure 6: OTU_1 and OTU_3 activity dynamics along TWS degradation representative points. 16S rRNA relative abundance is detailed for the two replicates ST1 (full colors) and ST2 (shaded colors). Curves corresponding to OTU dynamics were drawn using average replicated values.



V.4. Discussion

Two different enrichments were realized using *N. ephratae* gut microbiome as inocula, and sterile and non-sterilized straw as substrate. The enrichment on non-sterilized straw displayed decreasing LC degradation capacities whereas the second, with sterilized-straw, reached a stable degradation of 37%. Diversity analysis suggested that these different behaviors could result from the differences in community composition observed in both enrichments. Indeed, in both enrichments OTUs related to *Bacteroides* and *Prevotella* displayed similar abundance levels but when non-sterile straw was used, OTUs belonging to *Butyrivibrio* genus presented higher levels whereas the abundance of unclassified *Lachnospiraceae* and *Rikenellaceae RC9* decreased compared with the enrichment performed on sterilized straw. These last two families were dominated by unique OTUs, respectively OTU_3 and OTU_2 which were unfortunately not well assigned. Nevertheless, while assigned as a *Lachnospiraceae*, OTU_3 presented 100% identity to a novel *Clostridia* (GenBank: LN868251.1; unpublished), currently classified as a *Clostridium*, that was isolated from a biogas reactor and was described as a cellulose degrader. OTU_2 presented 100% identity with uncultured fiber-associated, probably lignocellulolytic active, rumen bacteria (Zened et al., 2012). OTU_1, the most dominant OTU, was related to *Bacteroides*, a genus known for its xylanolytic and hemicellulolytic activity (Tomomi Nishiyama, 2009). Despite these composition variations, the final communities showed strong similarities, being mainly

constituted by *Bacteroidetes* (mainly genus *Bacteroides*) and *Firmicutes* (mainly family *Clostridiaceae*) with both termite enrichments (irrespective of the substrate sterility). Similar compositions at the phylum level were also observed on a previous cow rumen-derived enrichment.

Members of the *Bacteroidetes* phylum are extremely rare in termite gut, so the selection of *Bacteroides* and *Prevotella* in the termite-derived enrichment, as well as the presence of other OTUs common to TWS and RWS but representing a minor fraction of the initial termite gut community is not explained. Cross-contamination by the lab environment cannot be excluded as the termite and rumen reactors were conducted in the same lab under aseptic but not sterile conditions. However, it is possible to remark that the first strong diversity switch observed between the initial termite gut community and the *Firmicutes*-dominated community at the end of the first culture cycle (Figure 1, C1) occurred in strictly sterile conditions. A second diversity switch was again observed during Cycle 2 that displayed a strong dominance of *Bacteroidetes* related OTUs at the expenses of the *Firmicutes* species originally presents in the termite inoculum. *Bacteroides* and *Prevotella* are known as lignocellulose degraders presenting high growth rates, so even if they were not abundant in the initial inocula, they were the main groups selected under the fermentation conditions applied in this study. Indeed, parameters such as pH and temperature were strongly different between the termite gut (neutral pH, few information about temperature (Brune, 2014)) and the lab reactors (pH 6.15 and 35°C). It could be expected that other parameters such as the nutrient composition or the concentration of metabolites potentially present in the termite gut differs from the conditions present in the bioreactor. It is possible that *Bacteroidetes* related OTUs selected in the reactors were more adapted to grow on wheat straw with the applied conditions. The selection of similar microbial communities from different inocula growing on the same substrate suggests that substrate and environmental conditions were the main drivers of the microbial selection. These results are in agreement with previous studies that reported a strong "substrate effect" on the microbial community composition (Eichorst et al., 2014; Simmons et al., 2014) after selection using different sources of inocula but the same substrate. However, despite the presence of common OTUs between RWS and TWS and the composition similarity, the community structure showed some clear differences specific to each inoculum source. The four main genii of TWS represented up to 95% of the community, so its diversity was much lower compared to RWS. Statistical analysis and ANOVA confirmed that the main differences between RWS and TWS were the difference abundances of some OTUs shared

between the two systems, and the presence of various OTUs specific to RWS, a community which was more rich and balanced. As previously observed with RWS (Auer et al., in preparation), the strongest shift in the profile of metabolically activity bacteria occurred during the LC degradation peak, so the initial or final points, generally characterized in most of previous studies, may actually not reflect the real active community.

TWS displayed high LC degradation capacity despite its lower diversity. Due to lower diversity compared to RWS and others described lignocellulolytic communities, it was here easier to identify potential correlations between degradation parameters and given OTUs. 16S rDNA profiles showed a lower variability than rRNA ones, with a strong dominance of a OTU related to *Bacteroides*. Interpreting 16S rRNA relative abundance as an indicator of microbial activity has some limitations (Blazewicz et al., 2013), so the differences in relative activities between taxa may not directly reflect the actual differences between them. However, even if absolute differences cannot be affirmed, 16S rRNA dynamics reflect changes in the activity of the different microbial groups present in the ecosystem; the observed switches in 16S rRNA profiles are good indicators of the community functioning. While 16S rDNA showed an early increase of minor genus (mainly *Proteobacteria*), these were quickly overtaken by the dominant genus (*Bacteroides* and *Clostridium*) since day two. This initial quick growth was probably due to the degradation of the more easily accessible components of LC, such as soluble sugars or proteins present in wheat straw. 16S rRNA data showed that this initial "blew" is probably very short and was not so clear considering bacterial activity. Moreover, it was followed since day one by a *Bacteroides* activity peak, itself followed by a *Lachnospiraceae* peak, whereas their rDNA relative abundance remained much more stable. Considering 16S rRNA levels, OTU_1 (*Bacteroides*) and OTU_3 (*Lachnospiraceae* but potential *Clostridium*) were extremely active, whereas OTU_2 *Rikenellaceae* had a very low activity, despite its non-negligible 16S rDNA abundance. *Bacteroides*-OTU_1 activity peak was well correlated with the peak of LC degradation rate whereas *Lachnospiraceae*-OTU3 peaked one or two days later. These results are in accordance with previous studies showing that non-adherent *Bacteroides* outmatched *Lachnospiraceae* on easy substrates, but on more recalcitrant substrate were then themselves outmatched by adherent *Lachnospiraceae* bacteria, with a slower growth but better degradation capacities (Macfarlane, 2006; Biddle et al., 2013).

V.5. Conclusion

The enrichment of termite gut bacteria on wheat straw resulted in a community TWS that was completely different from the initial termite gut community but presented good LC degradation capacities. This diversity switch was not due to the straw-resident community because the use of sterilized straw lead to the same type of profile. The enriched community in TWS was rather similar to RWS, a rumen-derived enrichment obtained with the same substrate and culture conditions. This result seems to correspond to the "substrate effect", observed in previous studies using different sources of inoculum and substrates. The enrichment on wheat straw as sole carbon source, inoculated with cow rumen or termite gut communities, resulted in the selection of communities dominated mainly by *Bacteroides* and *Firmicutes* (*Lachnospiraceae* and *Clostridium*), but with strong differences in community diversity and structure. Based on 16S rRNA abundances, the metabolically active bacteria shown a greater variability than 16S rDNA profiles, and the dynamics of some taxa appeared to be well correlated with LC degradation, highlighting potential functional interplays for LC hydrolysis between *Bacteroides* and *Firmicutes* members.

V.6. Acknowledgments

This research was supported by the French National Institute for Agronomical Research (INRA) and the Region Languedoc-Roussillon Midi-Pyrénées. This study was realized in collaboration with the Institute for Research and Development (IRD). The authors thank the Genomics and Transcripmics (GeT) INRA platform for their help with sequencing. We acknowledge Alain Robert and Isabel Monteiro for their help with the maintenance of termite nests, and Gunnar Oelker for its assistance in the experimental work.

V.7. References

Agler, M.T., Wrenn, B.A., Zinder, S.H., and Angenent, L.T. (2011). Waste to bioproduct conversion with undefined mixed cultures: the carboxylate platform. Trends Biotechnol. *29*, 70–78.

Biddle, A., Stewart, L., Blanchard, J., and Leschine, S. (2013). Untangling the Genetic Basis of Fibrolytic Specialization by Lachnospiraceae and Ruminococcaceae in Diverse Gut Communities. Diversity *5*, 627–640.

Blazewicz, S.J., Barnard, R.L., Daly, R.A., and Firestone, M.K. (2013). Evaluating rRNA as an indicator of microbial activity in environmental communities: limitations and uses. ISME J. *7*, 2061–2068.

Brune, A. (2014). Symbiotic digestion of lignocellulose in termite guts. Nat. Rev. Microbiol. *12*, 168–180.

Cao, K.-A.L., González, I., and Déjean, S. (2009). integrOmics: an R package to unravel relationships between two omics datasets. Bioinformatics *25*, 2855–2856.

Eichorst, S.A., Joshua, C., Sathitsuksanoh, N., Singh, S., Simmons, B.A., and Singer, S.W. (2014). Substrate-Specific Development of Thermophilic Bacterial Consortia by Using Chemically Pretreated Switchgrass. Appl. Environ. Microbiol. *80*, 7423–7432.

Feng, Y., Yu, Y., Wang, X., Qu, Y., Li, D., He, W., and Kim, B.H. (2011). Degradation of raw corn stover powder (RCSP) by an enriched microbial consortium and its community structure. Bioresour. Technol. *102*, 742–747.

Gao, Z.-M., Xu, X., and Ruan, L.-W. (2013). Enrichment and characterization of an anaerobic cellulolytic microbial consortium SQD-1.1 from mangrove soil. Appl. Microbiol. Biotechnol. *98*, 465–474.

Hongoh, Y., Deevong, P., Hattori, S., Inoue, T., Noda, S., Noparatnaraporn, N., Kudo, T., and Ohkuma, M. (2006). Phylogenetic diversity, localization, and cell morphologies of members of the candidate phylum TG3 and a subphylum in the phylum Fibrobacteres, recently discovered bacterial groups dominant in termite guts. Appl. Environ. Microbiol. *72*, 6780–6788.

Klein-Marcuschamer, D., Oleskowicz-Popiel, P., Simmons, B.A., and Blanch, H.W. (2012). The challenge of enzyme cost in the production of lignocellulosic biofuels. Biotechnol. Bioeng. *109*, 1083–1087.

Lazuka, A., Auer, L., Bozonnet, S., Morgavi, D.P., O'Donohue, M., and Hernandez-Raquet, G. (2015). Efficient anaerobic transformation of raw wheat straw by a robust cow rumen-derived microbial consortium. Bioresour. Technol. *196*, 241–249.

Macfarlane, S., and Macfarlane, G.T. (2006). Composition and metabolic activities of bacterial biofilms colonizing food residues in the human gut. Appl. Environ. Microbiol. *72*, 6204–6211.

Magoč, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinforma. Oxf. Engl. *27*, 2957–2963.

McMurdie, P.J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PloS One *8*, e61217.

Menon, V., and Rao, M. (2012). Trends in bioconversion of lignocellulose: Biofuels, platform chemicals & biorefinery concept. Prog. Energy Combust. Sci. *38*, 522–550.

Mikaelyan, A., Köhler, T., Lampert, N., Rohland, J., Boga, H., Meuser, K., and Brune, A. (2015). Classifying the bacterial gut microbiota of termites and cockroaches: A curated phylogenetic reference database (DictDb). Syst. Appl. Microbiol. *38*, 472–482.

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., et al. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl. Environ. Microbiol. *75*, 7537–7541.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., et al. (2014). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539–539.

Simmons, C. w., Reddy, A. p., Simmons, B. a., Singer, S. w., and VanderGheynst, J. s. (2014). Effect of inoculum source on the enrichment of microbial communities on two lignocellulosic bioenergy crops under thermophilic and high-solids conditions. J. Appl. Microbiol. *117*, 1025–1034.

Stamatakis, A. (2014). RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics btu033.

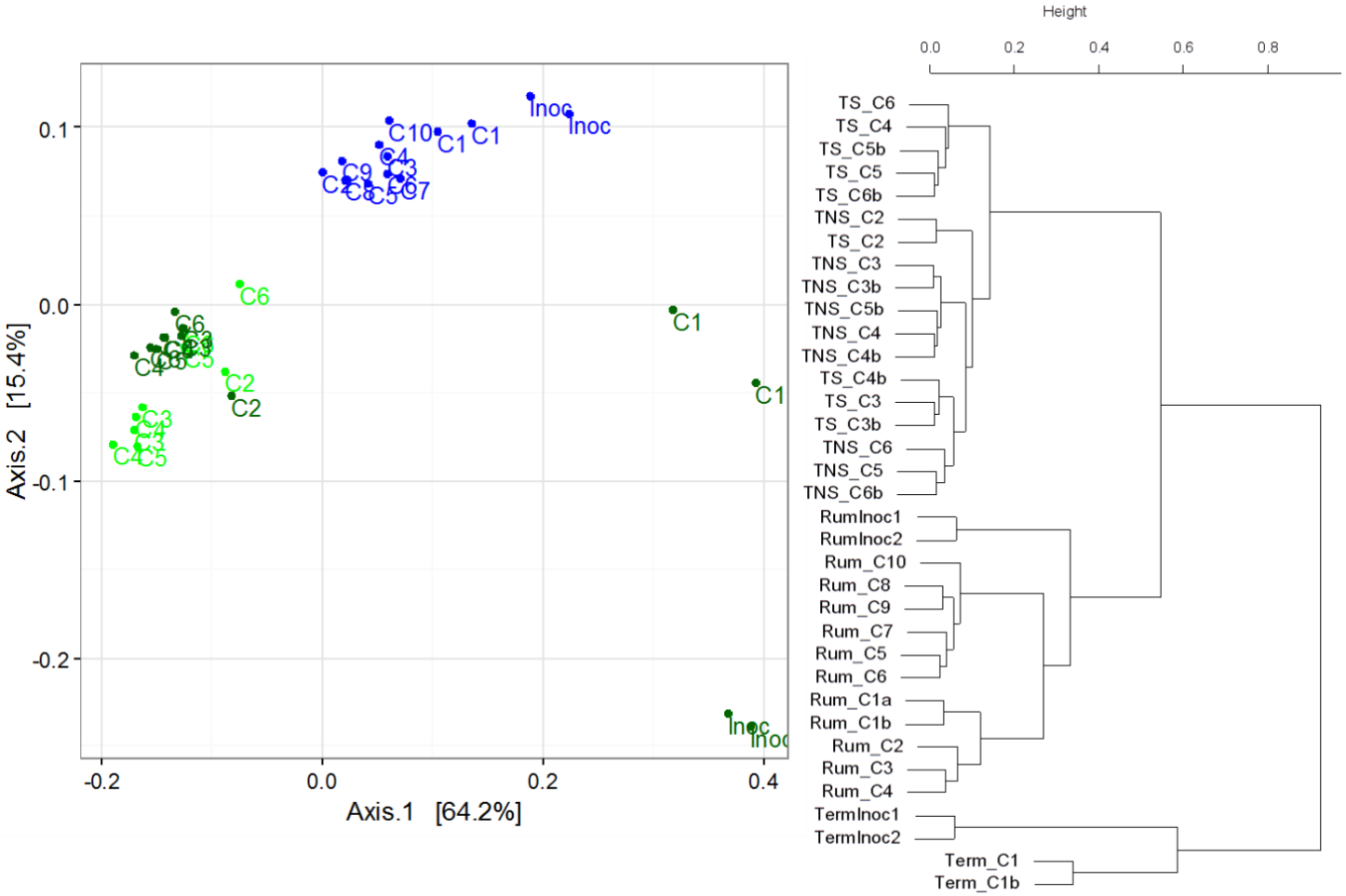
Tomomi Nishiyama, A.U. (2009). Bacteroides graminisolvens sp. nov., a xylanolytic anaerobe isolated from a methanogenic reactor treating cattle waste. Int. J. Syst. Evol. Microbiol. *59*, 1901–1907.

Yarza, P., Richter, M., Peplies, J., Euzeby, J., Amann, R., Schleifer, K.-H., Ludwig, W., Glöckner, F.O., and Rosselló-Móra, R. (2008). The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst. Appl. Microbiol. *31*, 241–250.

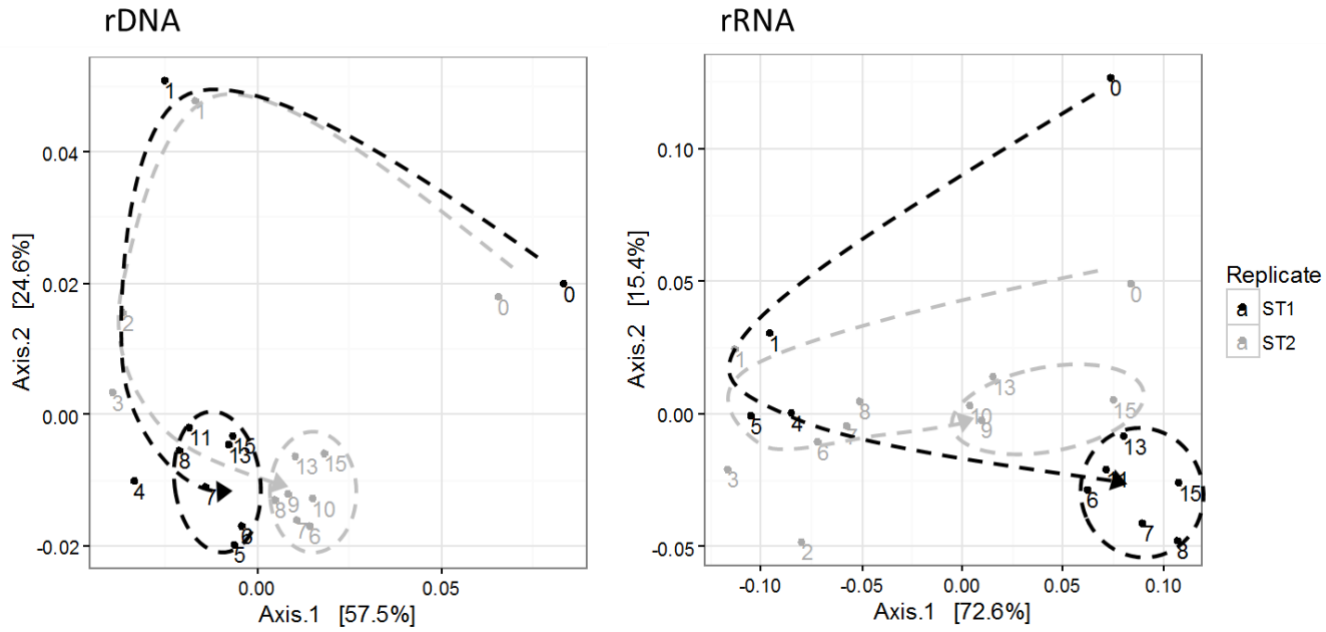
Zened, A., Combes, S., Cauquil, L., Mariette, J., Klopp, C., Bouchez, O., Troegeler-Meynadier, A., and Enjalbert, F. (2012). Microbial ecology of the rumen evaluated by 454 GS FLX pyrosequencing is affected by starch and oil supplementation of diets. FEMS Microbiol. Ecol. n/a–n/a.

V.8. Supplementary data

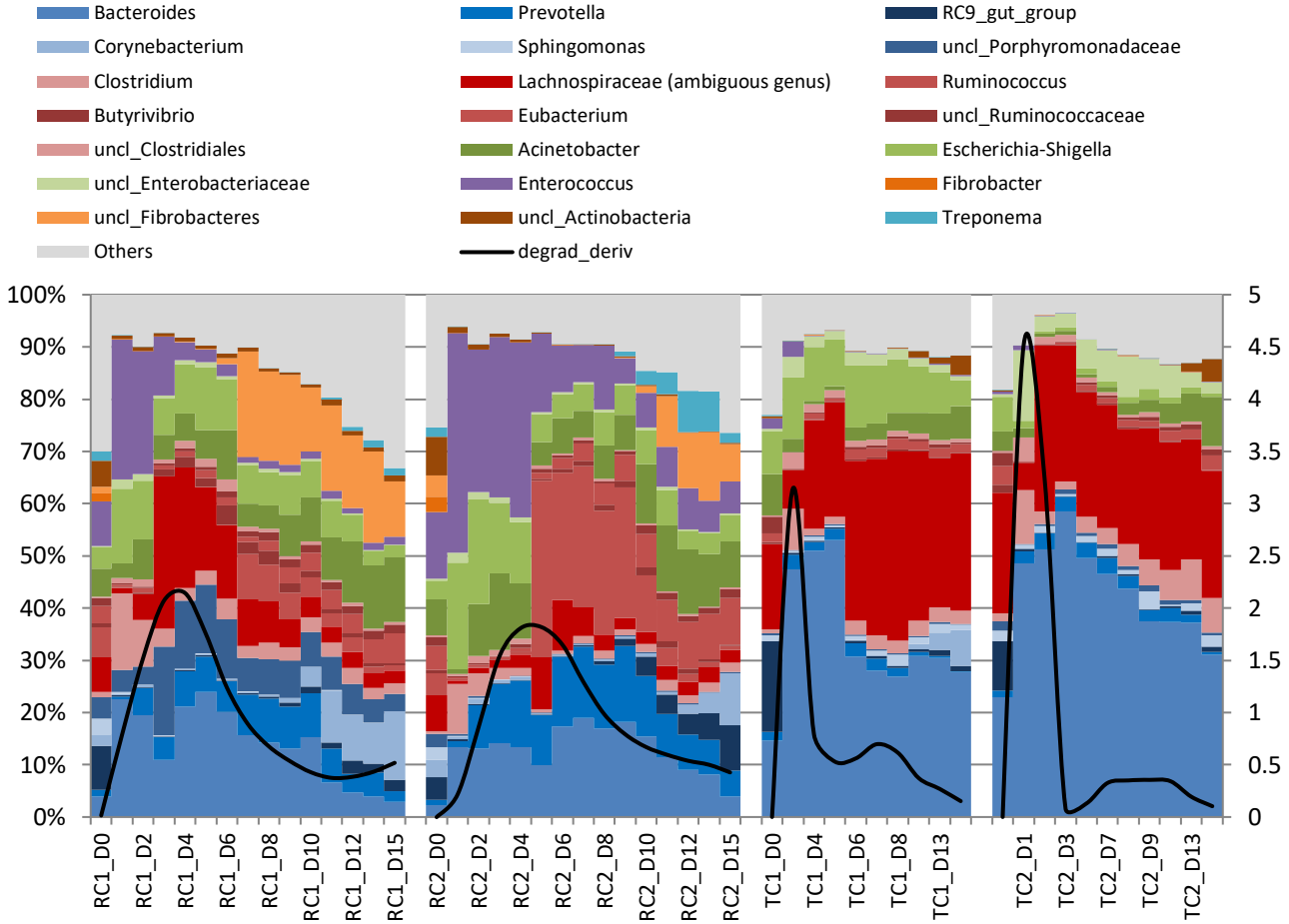
Supplementary data 1: Principle Coordinate Analysis based in weighted-Unifrac distances and clustering. RWS (blue, noted Rum) and TWS samples with sterile (dark green, noted TS) and non-sterile straw (light green, noted TNS) are detailed with their corresponding enrichment cycle, noted Ci. TWS initial termite gut inoculum and first cycle appeared to be very distant from all other samples, which clustered together close to RWS samples. Samples with sterile or non-sterile straw generally grouped together, but there were only small distances between them.



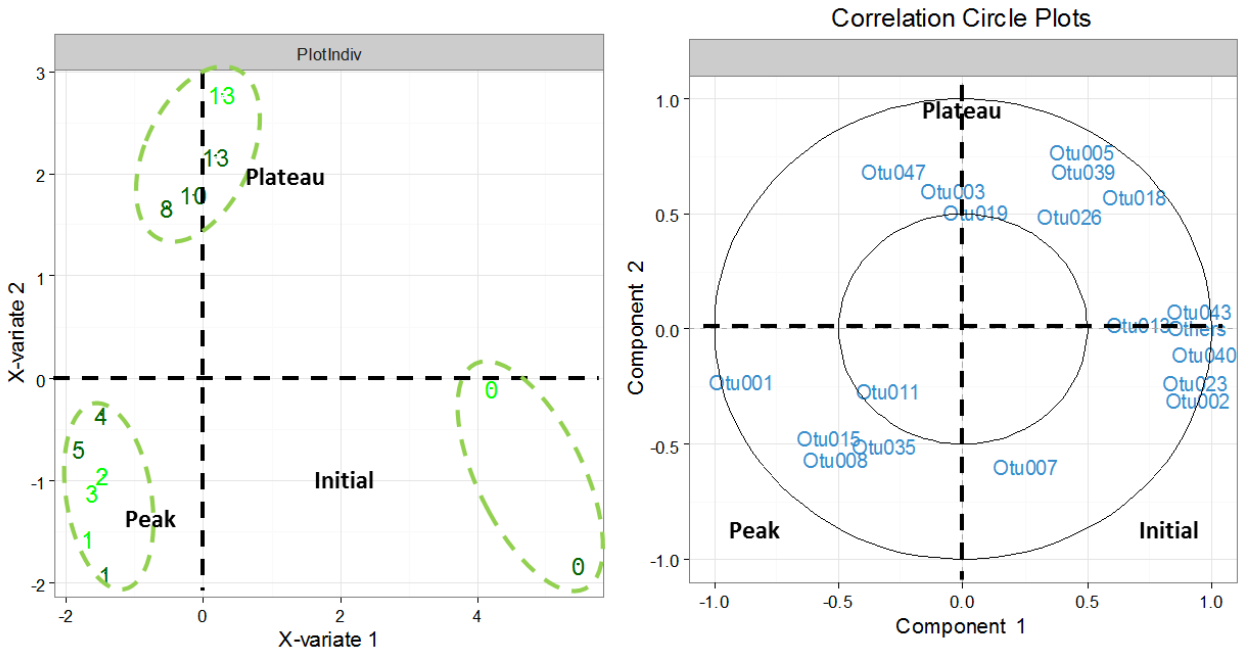
Supplementary data 2: Principle Coordinate Analysis (PCoA) of TWS community dynamics, based on weighted-Unifrac distances. Numbers correspond to sampling days, colors to replicates. With 16S rDNA data, communities stabilize around a final state after 5 to 6 days. Stabilization is slower with 16S rRNA data, especially for ST2 replicate where it took 9 days. With both replicates and both data type, distances were the strongest between early and final communities.



Supplementary data 3: RWS (RC1 and RC2) and TWS (TC1 and TC2) rRNA relative abundance profiles dynamics during biologically replicated LC degradation cycles. Minor genus are regrouped in "Others", whereas more dominant genus are colored according to their phylum or class (*Bacteroidetes* in blue, *Firmicutes/Clostridia* in red, *Firmicutes/Bacilli* in purple, *Proteobacteria* in green, *Fibrobacteres* in orange and *Spirochaetes* in cyan). The LC degradation rate (in $g.L^{-1}.day^{-1}$) is drawn with a black line.



Supplementary data 4: sPLS-DA analysis based on 16S rDNA data of TWS representative points (1 for initial, 3 for degradation peak and 2 for plateau, for each replicate), using degradation phases as discriminant factor. On the left, sPLS-DA TWS communities, labelled by time (in days). The corresponding correlation circle is presented on the right. OTUs average rRNA relative abundances during each phases is detailed in the table below, with taxonomic information and statistical significance evaluated with an ANOVA approach.



			Average rRNA relative abundance			ANOVA	
OTU	Class	Taxonomy	Initial	Peak	Plateau	P-value	
1	*Bacteroidia*	*Bacteroides*	13,63%	43,30%	28,99%	1,31E-04	***
2	*Bacteroidia*	*Rikenellaceae RC9*	10,87%	0,22%	0,12%	6,881E-08	***
3	*Clostridia*	*uncl Lachnospiraceae*	18,65%	18,57%	27,63%	0,2803	
5	*γ-proteobacteria*	*Acinetobacter*	2,78%	0,76%	3,76%	0,0001397	***
7	*Bacilli*	*uncl Lactobacillales*	1,23%	0,74%	0,12%	0,3781	
8	*Bacteroidia*	*Prevotella*	0,58%	1,58%	0,75%	0,05992	.
11	*Bacteroidia*	*Prevotella*	0,07%	0,17%	0,06%	0,3199	
13	*Clostridia*	*Butyrivibrio*	0,93%	0,10%	0,21%	0,01029	*
15	*Clostridia*	*uncl Clostridiales*	0,07%	2,06%	0,66%	0,09323	.
18	*Clostridia*	*Eubacterium*	2,59%	0,42%	1,93%	0,001273	**
19	*Actinobacteria*	*Corynebacterium*	0,08%	0,05%	0,86%	0,4201	
23	*Spirochaetes*	*Treponema*	0,19%	0,01%	0,02%	0,0003238	***
26	*α-proteobacteria*	*Sphingomonas*	1,46%	0,61%	1,33%	0,1191	
35	*Clostridia*	*Clostridium*	0,01%	0,11%	0,01%	0,336	
39	*β-proteobacteria*	*Comamonadaceae*	1,04%	0,15%	1,23%	0,0304	*
40	*α-proteobacteria*	*Rhodobacter*	0,84%	0,02%	0,14%	1,33E-07	***
43	*Spirochaetes*	*uncl Spirochaetales*	0,27%	0,00%	0,07%	0,0001491	***
47	*α-proteobacteria*	*Pleomorphomonas*	0,09%	0,30%	0,71%	0,01941	*
		Others	26,59%	4,97%	11,44%	0,0004335	***

VI. Conclusion du chapitre

La procédure d'enrichissement sur substrat stérile a permis de maintenir les capacités de dégradation de l'inoculum, la dégradation en fin de cycle ayant à peine diminué pour se stabiliser autour de 38%. Cependant, malgré ce maintien apparent de la fonction de dégradation, on obtient en quelques cycles une communauté largement dominée par les *Bacteroidetes*, différente de la communauté initiale (*Spirochaetes*, *Fibrobacteres* et *TG3*) mais également de celle du premier cycle de culture décrit dans le chapitre VI (*Firmicutes*). La communauté obtenue, appelée TWS, est très simple et n'est composée que de quelques OTUs très majoritaires. Cependant, si les *Bacteroidetes* semblent largement majoritaires, cette observation est à nuancer quand on s'intéresse à l'activité des taxons, évaluée par séquençage de l'ARNr 16S. Les *Firmicutes* apparaissent particulièrement actif dans la communauté.

Au cours d'un cycle de dégradation, TWS présente une vitesse de dégradation de la lignocellulose plus élevée que RWS, avec un pic de dégradation plus fort et précoce. Sa diversité est également beaucoup plus basse, la communauté étant beaucoup plus simple, et l'activité bactérienne est principalement due aux mêmes phylums (*Bacteroidetes* et *Firmicutes*) qu'avec RWS. S'il est donc possible de stabiliser un inoculum lignocellulolytique en partant du microbiote intestinal de termite, la communauté finale obtenue ne ressemble pas du tout à l'initiale. L'effet « substrat » ainsi que les conditions de cultures imposées apparaissent donc être plus forts que l'effet « inoculum », même quand celui-ci est maximisé en comparant deux communautés initiales aux compositions très différent.

CONCLUSION ET PERSPECTIVES

CONCLUSION ET PERSPECTIVES

La lignocellulose est un composé végétal complexe constitué de cellulose, hémicelluloses et lignine. Composant majoritaire des parois végétales et abondamment présente dans les parties non alimentaires des plantes, elle est le biopolymère le plus abondant sur Terre mais est généralement considéré comme un déchet. Pourtant, dans la nature, elle représente une ressource alimentaire pour de nombreux organismes herbivores, des termites aux ruminants. Il est donc possible de valoriser la lignocellulose en tant que source de carbone, renouvelable qui plus est. Différentes approches sont classiquement envisagées, de l'utilisation de traitements physico-chimiques à celle de micro-organismes en culture pure en passant par la digestion enzymatique, avec différentes produits cibles, dont le méthane ou les biocarburants. Cependant, ces approches atteignent rapidement leurs limites, et l'étude des écosystèmes digestifs lignocellulolytiques naturels suggère une approche alternative : l'utilisation de communautés microbiennes anaérobies pour la bioconversion de lignocellulose en carboxylates. Ce travail de thèse a donc pour objectif d'étudier les communautés bactériennes dans un contexte particulier, celui de la valorisation d'une ressource lignocellulosique, ici de la paille de blé comme substrat modèle. L'enjeu de cette thèse est double : le premier, plus appliqué, est d'obtenir à partir de microbiotes intestinaux (rumen bovin et intestin de termite) des communautés bactériennes stables, actives sur des substrats lignocellulosiques, et productrices de carboxylates. De tels inocula n'ont jamais été utilisés dans ce contexte, et présentent une diversité bactérienne très différente des communautés actuellement décrites en fermenteur. La comparaison de leurs performances et de leur composition, entre elles ou avec celles rapportées dans les travaux d'autres équipes, permet donc également de mieux discuter de « l'effet inoculum » sur les communautés et sur la dégradation de la lignocellulose, en apportant des données obtenues en utilisant des inocula plus variées que ceux habituellement utilisés. Le deuxième enjeu, plus exploratoire, est d'améliorer notre compréhension du fonctionnement de ces communautés, ce qui est la première étape vers leur contrôle et leur maîtrise dans des procédés industriels comme la plateforme des carboxylates.

La première étape de ce travail de thèse a consisté à appliquer des outils d'écologie microbienne à nos systèmes, et à adapter les techniques d'analyse les plus récentes à nos problématiques. Dans cette étude nous avons utilisé comme marqueur de la diversité

microbienne le gène codant pour l'ARNr 16S ainsi que ses transcrits, qui ont été séquencés. Alors que les progrès des techniques de séquençage de nouvelle génération (NGS) offrent un accès de plus en plus facile à des données d'un volume de plus en plus grand, celui-ci peut devenir trop important pour les outils de traitement existants si le nombre d'échantillons à comparer devient élevé. Avec plus de trois cent échantillons séquencés à analyser simultanément, cette étude requérait d'adapter les méthodes d'analyse des donnés NGS disponibles. La méthode proposée et validée dans le premier chapitre est l'utilisation d'un filtre des séquences singletons sur les données brutes. Elle permet de réduire considérablement la complexité des données dès les premières étapes de traitement, et ainsi de diminuer drastiquement les besoins en mémoire et puissance et temps de calcul. De plus, comme cela a été vérifié à l'aide de données simulées, mais également de données de séquençage de communautés synthétiques et de communautés réelles, son impact est minime sur la description de la diversité des communautés. L'expertise acquise sur les techniques de séquençage et méthodes d'analyse a également servi à participer au développement de FROGS, une solution innovante de traitement de données de séquençage, de plus entièrement disponible à travers une interface web Galaxy.

La méthode d'analyse validée précédemment a donc pu être utilisée pour étudier les communautés microbiennes obtenues par inoculation de rumen bovin, une communauté microbienne très étudiée mais jamais utilisée dans le contexte de la plateforme des carboxylates. Une communauté stable et active sur lignocellulose a été obtenue après 10 cycles de culture. Nommée RWS, elle permet de dégrader jusqu'à 55% du substrat (paille de blé) en 15 jours, remplissant ainsi un des objectifs de la thèse. RWS est composée majoritairement de *Bacteroidetes* et *Firmicutes*, une composition similaire à la fois au rumen initial et aux communautés obtenues dans d'autres études. Cependant, l'analyse à des niveaux taxonomiques plus précis montre la présence de nombreux genres jusqu'ici rarement observés en fermenteurs, mais courants dans le rumen (e.g. *Prevotella, Bacteroides, Ruminoccocus*…). En plus d'être une des communautés enrichies les plus actives sur lignocellulose actuellement décrites, RWS présente des niveaux élevés d'activité enzymatiques (notamment xylanase) et une diversité inédite.

L'utilisation de prétraitements physico-chimiques afin d'améliorer la dégradabilité des substrats est quasiment incontournable dans les procédés actuels de valorisation, mais ceux-ci sont variés et leur effet sur les communautés microbiennes est très peu connu. L'étude de l'impact du prétraitement du substrat par des méthodes de traitement chimique par voie sèche

(imprégnation à la soude) a montré que toutes les populations bactériennes qui composent RWS n'y sont pas sensibles de la même manière. Beaucoup d'entre elles sont affectées négativement, alors que d'autres sont favorisées et leur rôle est augmenté dans la dégradation. En effet, les prétraitements conduisent à une augmentation de la vitesse de dégradation, et à une augmentation de la dégradation maximale pour l'un d'entre eux (atteignant plus de 60%). Cette étude a aussi servi de cadre à une caractérisation dynamique des populations bactériennes et de leur activité par séquençage de l'ADNr 16S et de l'ARNr 16S. Celle-ci a montré un comportement cyclique des communautés, les différentes phases de la dégradation (initial, pic, final) étant associées à des profils de population et d'activités différents : on observe des pics d'activité de certaines populations, principalement affiliées aux classes *Bacteroidia* et *Clostridia*, alors que l'état final est caractérisé par une diversité plus grande, les groupes majoritaires étant moins dominants, et les minoritaires plus représentés. L'observation à l'état final, classiquement réalisée, n'est donc pas la plus représentative de l'état de la communauté pendant le pic de dégradation. Même si l'effet des prétraitements sur les communautés s'est révélé statistiquement plus faible que leur évolution au cours de la dégradation, l'étude de l'effet des prétraitements a par ailleurs permis l'identification de populations dont l'activité est corrélée à un type de prétraitement. Ainsi, *Bacteroides* et *Clostridium* sont majoritaires pour tous les traitements mais leur activité est particulièrement exacerbée par les prétraitements à la soude.

Le microbiote intestinal de termite est très différent des communautés ruminales et des communautés lignocellulolytiques décrites en fermenteur. Il varie considérablement en fonction des espèces des termites et de leur régime alimentaire, mais présente de manière générale très peu de *Bacteroidetes* et *Firmicutes*, étant plutôt dominé par *Spirochaetes* et *Fibrobacteres* entre autres. Nous avons donc entrepris d'étudier le potentiel du microbiome intestinal de quatre espèces de termites (*Nasutitermes ephratae* et *N. lujae*, *Microcerotermes parvus* et *Termes hospes*) à transformer la paille de blé en conditions contrôlées en réacteur. L'inoculation par du microbiote intestinal de termite a donné des résultats satisfaisants en termes de dégradation, notamment avec *Nasutitermes ephratae*. Celui-ci a permis d'atteindre des pourcentages de dégradation de 40% dès le premier cycle de 15 jours de culture, soit une bien meilleure dégradation que celle obtenue avec du rumen au même stade (environ 20% en 8 jours). L'inoculation d'intestins de termites semble donc être très prometteuse en vue d'obtenir une communauté extrêmement active sur lignocellulose. Cependant, du point de vue de la diversité bactérienne, la composition initiale des communautés n'a pas été maintenue et

a basculé vers une dominance très forte des *Firmicutes*. Avec les quatre espèces de termites testées, les *Spirochaetes, Fibrobacteres* et *TG3* initialement dominants ont laissé la place à des *Firmicutes* (notamment *Clostridium termitidis*, des *Lachnospiraceae* et des assimilés *Ruminococcaceae*) et des *Proteobacteria* (*Enterobacteriaceae* pour la plupart), parfois accompagnées par des *Bacteroidetes* (*Dysgonomonas*). Par ailleurs, la diversité bactérienne finale s'est révélée très peu répétable entre réplicats biologiques, donc encore potentiellement assez éloignée de son état stable.

Il était donc nécessaire de stabiliser ces communautés issues de termites. Un enrichissement par repiquages successifs permet cette stabilisation, et s'était dans le cas du rumen bovin accompagné d'une augmentation des capacités de dégradation. Ce procédé a donc été appliqué à la communauté issue du microbiote intestinal de *Nasutitermes ephratae*, qui avait donné les meilleurs résultats lors du premier cycle de culture. La stabilisation par enrichissement du microbiote de *N. ephratae* n'a malheureusement pas conduit à une augmentation aussi forte qu'avec le rumen bovin de ses capacités de dégradation. La communauté obtenue, TWS, atteint environ 50% de dégradation, et est donc légèrement moins efficace que RWS. Sa composition présente des similarités fortes avec RWS : sa diversité active est beaucoup plus faible, mais comme pour RWS, est principalement partagée entre genre *Bacteroides* (phylum *Bacteroidetes*) et classe *Clostridia* (*Firmicutes*). Sa faible diversité métaboliquement active (3 OTUs seulement comptant pour 90% des ARNr 16S au moment du pic de dégradation) fait de TWS un bon sujet d'étude, sa complexité assez basse rendant plus facile l'identification de corrélations avec des paramètres cinétiques de la dégradation. La composition de TWS a également permis d'établir la prédominance de « l'effet substrat » sur « l'effet inoculum », puisque malgré des compositions initiales extrêmement différentes, les conditions de culture et le substrat ont amené à la sélection de communautés bien plus proches qu'elles ne l'étaient initialement.

Pris tous ensemble, ces travaux ont donc montré que les écosystèmes digestifs, dont l'importance comme source de nouvelles enzymes pour la plateforme des sucres n'est plus à démontrer, sont également de très bons inocula pour la plateforme carboxylates. Dans ce contexte, les conditions de culture (dont le choix du substrat) ont cependant un effet extrêmement fort et façonnent les communautés sélectionnées qui peuvent être très différentes des communautés initialement inoculées. Quelles que soient les études, la plateforme carboxylate semble favoriser très fortement les *Bacteroidetes* et *Firmicutes*, qui sont également les phyla les plus actifs pendant les pics de dégradation, où ils ne sont représentés

que par quelques OTUs très majoritaires. Si une diversité élevée est primordiale pour permettre la bonne adaptabilité d'un système, la dégradation semble par contre être réalisée en majeure partie par une plus petite partie de la communauté.

Afin de confirmer avec certitude le rôle d'une population au sein de la communauté, l'approche classique serait de réaliser des expériences de dégradation de lignocellulose par des communautés synthétiques. Il serait donc intéressant de cultiver les quelques OTUs majoritaires, isolément ou en co-cultures, afin d'étudier dans un second temps les performances de différents mélanges, et ainsi identifier les fonctions de chaque membre de la communauté. L'approche « procédés » serait, en conservant la communauté indivisée, de modifier les conditions de culture, notamment en ajoutant des substrats spécifiques (cellulose, hémicellulose, protéines…). L'effet de ces ajouts sur chaque population de la communauté, suivi par séquençage de la diversité (et éventuellement d'autres méthodes -omiques) permettrait d'établir des utilisations préférentielles de substrat, et donc la fonction de chaque population dans l'ensemble de la communauté. Cette approche permettrait par ailleurs de conserver les populations minoritaires, et d'étudier leur rôle dans le fonctionnement du système. Comprendre le fonctionnement global des communautés microbiennes lignocelluloytiques en fermenteur permettrait ainsi d'envisager des moyens de contrôle des différentes sous-populations, et donc de maîtrise de la communauté.

Si les effets les plus visibles sont ceux des populations dominantes de la communauté, il est cependant probable que les taxons minoritaires jouent tout de même un rôle. Ceux-ci sont actifs notamment en fin de cycle de dégradation. Deux cas de figures peuvent se présenter : soit ils sont un acteur secondaire de la dégradation, exploitant les carboxylates libérés ou la biomasse bactérienne en déclin pour leur développement, soit ils sont un acteur tardif dont l'activité n'est perceptible que sur la partie très réfractaire à la dégradation du substrat. La taxonomie peut être informative sur ce point, puisque certains taxons, comme les *Fibrobacteres* de RWS, ne sont actifs qu'après le pic de dégradation mais sont connus pour être cellulolytiques. Ceux-ci pourraient donc être liés à la dégradation plus lente qui persiste en fin de cycle, et apporter des activités enzymatiques capables d'attaquer un substrat réfractaire en prenant le relai des taxons dominants, capables de ne dégrader qu'une fraction du substrat. L'état libre ou attaché au substrat, parfois caractéristique de certaines bactéries (*Fibrobacteres* et *Clostridia* plutôt adhérents, *Bacteroidetes* plutôt libres) pourrait également être un indicateur de ses spécificités de dégradation. Le fractionnement d'échantillons de fermenteur entre bactéries libres et adhérentes et le séquençage séparé de ces fractions, réalisé

dans la poursuite du projet, permettra de confirmer une répartition spatiale spécifique à des populations, et d'étudier la dynamique du phénomène.

Enfin, afin de mieux déterminer le rôle de chaque population au sein de la communauté, plusieurs approches peuvent être envisagées pour dépasser le stade des hypothèses, et éventuellement aller en complément d'une approche par communautés synthétiques ou réponses à des perturbations. L'analyse du transcriptome à l'aide de puces à ADN porteuses de sondes spécifiques des « carbohydrate enzymes » (CAZy) est une possibilité, testée en collaboration avec une autre équipe du laboratoire. Ces expériences sont en cours d'analyse, mais si les résultats sont prometteurs, la technologie des puces à ADN a ses limites dans le cas de communautés microbiennes complexes : seuls les gènes portés par la puce sont visibles, et l'analyse est donc très dépendante du design de la puce, celui-ci lui-même dépendant de la base de données utilisée. La résolution taxonomique de la CAZy database étant faible, il sera difficile d'associer la détection d'une enzyme à l'activité d'un taxon donné. La deuxième approche est l'étude des protéines, incluant les enzymes d'intérêt, par métaprotéomique. Cette technique est actuellement utilisée pour la suite du projet dans l'équipe et commence à donner ses premiers résultats, permettant notamment de montrer une complémentarité fonctionnelle entre *Bacteroidetes* et *Firmicutes*, qui produisent chacun certaines familles d'enzymes spécifiques. Cependant, une assignation taxonomique fine est là encore difficile, et la technique est elle aussi très dépendante des bases de données. Enfin, une approche par métatranscriptomique *de novo* associée ou non à du séquençage de métagénome et à de la reconstruction de génome aurait les meilleurs résultats en termes d'assignation taxonomique, et n'est pas sensible aux bases de données. Cependant, elle est complexe et coûteuse à mettre en œuvre, et si elle est capable de « voir » des transcrits absents des banques de données, se pose alors le problème de leur annotation. Aucune de ces approches n'est à elle seule parfaite, mais elles peuvent ensembles aboutir à une compréhension fine du fonctionnement des écosystèmes bactériens en fermenteurs, étape essentielle vers la maîtrise des communautés microbiennes lignocellulolytiques.

Références bibliographiques

REFERENCES BIBLIOGRAPHIQUES

Aanen, D.K., Ros, V.I., de Fine Licht, H.H., Mitchell, J., de Beer, Z.W., Slippers, B., Rouland-LeFèvre, C., and Boomsma, J.J. (2007). Patterns of interaction specificity of fungus-growing termites and Termitomyces symbionts in South Africa. BMC Evol. Biol. *7*, 115.

Agbor, V.B., Cicek, N., Sparling, R., Berlin, A., and Levin, D.B. (2011). Biomass pretreatment: Fundamentals toward application. Biotechnol. Adv. *29*, 675–685.

Agler, M.T., Wrenn, B.A., Zinder, S.H., and Angenent, L.T. (2011). Waste to bioproduct conversion with undefined mixed cultures: the carboxylate platform. Trends Biotechnol. *29*, 70–78.

Alizadeh, H., Teymouri, F., Gilbert, T.I., and Dale, B.E. (2005). Pretreatment of switchgrass by ammonia fiber explosion (AFEX). Appl. Biochem. Biotechnol. *121–124*, 1133–1141.

Baldwin, B.G. (1992). Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: An example from the compositae. Mol. Phylogenet. Evol. *1*, 3–16.

Barakat, A., de Vries, H., and Rouau, X. (2013). Dry fractionation process as an important step in current and future lignocellulose biorefineries: A review. Bioresour. Technol. *134*, 362–373.

Barakat, A., Chuetor, S., Monlau, F., Solhy, A., and Rouau, X. (2014). Eco-friendly dry chemo-mechanical pretreatments of lignocellulosic biomass: Impact on energy and yield of the enzymatic hydrolysis. Appl. Energy *113*, 97–105.

Berry, D., Mahfoudh, K.B., Wagner, M., and Loy, A. (2011). Barcoded Primers Used in Multiplex Amplicon Pyrosequencing Bias Amplification. Appl. Environ. Microbiol. *77*, 7846–7849.

Bignell, D.E., and Eggleton, P. (2000). Termites in Ecosystems. In Termites: Evolution, Sociality, Symbioses, Ecology, T. Abe, D.E. Bignell, and M. Higashi, eds. (Springer Netherlands), pp. 363–387.

Bignell, D.E., Roisin, Y., and Lo, N. (2010). Biology of Termites: a Modern Synthesis (Springer Science & Business Media).

Breznak, J.A., and Brune, A. (1994). Role of Microorganisms in the Digestion of Lignocellulose by Termites. Annu. Rev. Entomol. *39*, 453–487.

Brulc, J.M., Antonopoulos, D.A., Miller, M.E.B., Wilson, M.K., Yannarell, A.C., Dinsdale, E.A., Edwards, R.E., Frank, E.D., Emerson, J.B., Wacklin, P., et al. (2009). Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. Proc. Natl. Acad. Sci. *106*, 1948–1953.

Brune, A. (2014). Symbiotic digestion of lignocellulose in termite guts. Nat. Rev. Microbiol. *12*, 168–180.

Cai, Y., and Sun, Y. (2011). ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. Nucleic Acids Res. *39*, e95.

Canam, T., Town, J., Iroba, K., Tabil, L., and Dumonceaux, T. (2013). Pretreatment of Lignocellulosic Biomass Using Microorganisms: Approaches, Advantages, and Limitations. In Sustainable Degradation of Lignocellulosic Biomass - Techniques, Applications and Commercialization, A. Chandel, ed. (InTech), p.

Cao, K.-A.L., González, I., and Déjean, S. (2009). integrOmics: an R package to unravel relationships between two omics datasets. Bioinformatics *25*, 2855–2856.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat. Methods *7*, 335–336.

Case, R.J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W.F., and Kjelleberg, S. (2007). Use of 16S rRNA and rpoB Genes as Molecular Markers for Microbial Ecology Studies. Appl. Environ. Microbiol. *73*, 278–288.

Chandra, R., Takeuchi, H., and Hasegawa, T. (2012). Methane production from lignocellulosic agricultural crop wastes: A review in context to second generation of biofuel production. Renew. Sustain. Energy Rev. *16*, 1462–1476.

Chao, A. (1987). Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. Biometrics *43*, 783–791.

Chen, C.-Y. (2014). DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present. Front. Microbiol. *5*, 305.

Chen, X.L., Wang, J.K., Wu, Y.M., and Liu, J.X. (2008). Effects of chemical treatments of rice straw on rumen fermentation characteristics, fibrolytic enzyme activities and populations of liquid- and solid-associated ruminal microbes in vitro. Anim. Feed Sci. Technol. *141*, 1–14.

Chen, Y.-C., Eisner, J.D., Kattar, M.M., Rassoulian-Barrett, S.L., Lafe, K., Bui, U., Limaye, A.P., and Cookson, B.T. (2001). Polymorphic Internal Transcribed Spacer Region 1 DNA Sequences Identify Medically Important Yeasts. J. Clin. Microbiol. *39*, 4042–4051.

Cheng, Y.F., Edwards, J.E., Allison, G.G., Zhu, W.-Y., and Theodorou, M.K. (2009). Diversity and activity of enriched ruminal cultures of anaerobic fungi and methanogens grown together on lignocellulose in consecutive batch culture. Bioresour. Technol. *100*, 4821–4828.

Clark, T.A., Murray, I.A., Morgan, R.D., Kislyuk, A.O., Spittle, K.E., Boitano, M., Fomenkov, A., Roberts, R.J., and Korlach, J. (2011). Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. Nucleic Acids Res. gkr1146.

Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R., and Tiedje, J.M. (2014). Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. *42*, D633–D642.

Collins, F.S., Morgan, M., and Patrinos, A. (2003). The Human Genome Project: Lessons from Large-Scale Biology. Science *300*, 286–290.

Combes, S., Fortun-Lamothe, L., Cauquil, L., and Gidenne, T. (2013). Engineering the rabbit digestive ecosystem to improve digestive health and efficacy. Animal *7*, 1429–1439.

Cragg, S.M., Beckham, G.T., Bruce, N.C., Bugg, T.D., Distel, D.L., Dupree, P., Etxabe, A.G., Goodell, B.S., Jellison, J., McGeehan, J.E., et al. (2015). Lignocellulose degradation mechanisms across the Tree of Life. Curr. Opin. Chem. Biol. *29*, 108–119.

Dashtban, M., Schraft, H., and Qin, W. (2009). Fungal bioconversion of lignocellulosic residues; opportunities & perspectives. Int. J. Biol. Sci. *5*, 578–595.

DeAngelis, K.M., Fortney, J.L., Borglin, S., Silver, W.L., Simmons, B.A., and Hazen, T.C. (2012). Anaerobic decomposition of switchgrass by tropical soil-derived feedstock-adapted consortia. mBio *3*.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G.L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. *72*, 5069–5072.

Diouf, M., Roy, V., Mora, P., Frechault, S., Lefebvre, T., Hervé, V., Rouland-Lefèvre, C., and Miambi, E. (2015). Profiling the Succession of Bacterial Communities throughout the Life Stages of a Higher Termite Nasutitermes arborum (Termitidae, Nasutitermitinae) Using 16S rRNA Gene Pyrosequencing. PLoS ONE *10*.

Donovan, S.E., Eggleton, P., and Bignell, D.E. (2001). Gut content analysis and a new feeding group classification of termites. Ecol. Entomol. *26*, 356–366.

Ebert, A., and Brune, A. (1997). Hydrogen Concentration Profiles at the Oxic-Anoxic Interface: a Microsensor Study of the Hindgut of the Wood-Feeding Lower Termite Reticulitermes flavipes (Kollar). Appl. Environ. Microbiol. *63*, 4039–4046.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics *26*, 2460–2461.

Edgar, R.C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat. Methods *10*, 996–998.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. Bioinformatics *27*, 2194–2200.

Eichorst, S.A., Varanasi, P., Stavila, V., Zemla, M., Auer, M., Singh, S., Simmons, B.A., and Singer, S.W. (2013). Community dynamics of cellulose-adapted thermophilic bacterial consortia. Environ. Microbiol. *15*, 2573–2587.

Eichorst, S.A., Joshua, C., Sathitsuksanoh, N., Singh, S., Simmons, B.A., and Singer, S.W. (2014). Substrate-Specific Development of Thermophilic Bacterial Consortia by Using Chemically Pretreated Switchgrass. Appl. Environ. Microbiol. *80*, 7423–7432.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. Science *323*, 133–138.

Faraco, V. (2013). Lignocellulose Conversion: Enzymatic and Microbial Tools for Bioethanol Production (Springer Science & Business Media).

Feng, Y., Yu, Y., Wang, X., Qu, Y., Li, D., He, W., and Kim, B.H. (2011). Degradation of raw corn stover powder (RCSP) by an enriched microbial consortium and its community structure. Bioresour. Technol. *102*, 742–747.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

Gao, D., and Tao, Y. (2012). Current molecular biologic techniques for characterizing environmental microbial community. Front. Environ. Sci. Eng. *6*, 82–97.

Gao, Z.-M., Xu, X., and Ruan, L.-W. (2013). Enrichment and characterization of an anaerobic cellulolytic microbial consortium SQD-1.1 from mangrove soil. Appl. Microbiol. Biotechnol. *98*, 465–474.

Ghodsi, M., Liu, B., and Pop, M. (2011). DNACLUST: accurate and efficient clustering of phylogenetic marker genes. BMC Bioinformatics *12*, 271.

Giovannoni, S.J., Britschgi, T.B., Moyer, C.L., and Field, K.G. (1990). Genetic diversity in Sargasso Sea bacterioplankton. Nature *345*, 60–63.

Gonzalez, J.M., Portillo, M.C., Belda-Ferre, P., and Mira, A. (2012). Amplification by PCR Artificially Reduces the Proportion of the Rare Biosphere in Microbial Communities. PLoS ONE *7*, e29973.

Gordon, G.L., and Phillips, M.W. (1998). The role of anaerobic gut fungi in ruminants. Nutr. Res. Rev. *11*, 133–168.

Granda, C.B., Holtzapple, M.T., Luce, G., Searcy, K., and Mamrosh, D.L. (2009). Carboxylate Platform: The MixAlco Process Part 2: Process Economics. Appl. Biochem. Biotechnol. *156*, 107–124.

Guo, P., Mochidzuki, K., Cheng, W., Zhou, M., Gao, H., Zheng, D., Wang, X., and Cui, Z. (2011). Effects of different pretreatment strategies on corn stalk acidogenic fermentation using a microbial consortium. Bioresour. Technol. *102*, 7526–7531.

Haas, B.J., Gevers, D., Earl, A.M., Feldgarden, M., Ward, D.V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S.K., Sodergren, E., et al. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Res. *21*, 494–504.

Haegeman, B., Hamelin, J., Moriarty, J., Neal, P., Dushoff, J., and Weitz, J.S. (2013). Robust estimation of microbial diversity in theory and in practice. ISME J. *7*, 1092–1101.

Hendriks, A.T.W.M., and Zeeman, G. (2009). Pretreatments to enhance the digestibility of lignocellulosic biomass. Bioresour. Technol. *100*, 10–18.

Hess, M., Sczyrba, A., Egan, R., Kim, T.-W., Chokhawala, H., Schroth, G., Luo, S., Clark, D.S., Chen, F., Zhang, T., et al. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science *331*, 463–467.

Holtzapple, M.T., and Granda, C.B. (2009). Carboxylate Platform: The MixAlco Process Part 1: Comparison of Three Biomass Conversion Platforms. Appl. Biochem. Biotechnol. *156*, 95–106.

Hongoh, Y. (2011). Toward the functional analysis of uncultivable, symbiotic microorganisms in the termite gut. Cell. Mol. Life Sci. CMLS *68*, 1311–1325.

Hugenholtz, P., Goebel, B.M., and Pace, N.R. (1998). Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity. J. Bacteriol. *180*, 4765–4774.

Hui, W., Jiajia, L., Yucai, L., Peng, G., Xiaofen, W., Kazuhiro, M., and Zongjun, C. (2013). Bioconversion of un-pretreated lignocellulosic materials by a microbial consortium XDC-2. Bioresour. Technol. *136*, 481–487.

Huse, S.M., Welch, D.M., Morrison, H.G., and Sogin, M.L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environ. Microbiol. *12*, 1889–1898.

Jami, E., and Mizrahi, I. (2012). Composition and Similarity of Bovine Rumen Microbiota across Individual Animals. PLoS ONE *7*, e33306.

Janssen, P.H. (2006). Identifying the Dominant Soil Bacterial Taxa in Libraries of 16S rRNA and 16S rRNA Genes. Appl. Environ. Microbiol. *72*, 1719–1728.

Janssen, P.H., and Kirs, M. (2008). Structure of the Archaeal Community of the Rumen. Appl. Environ. Microbiol. *74*, 3619–3625.

Jeoh, T., Ishizawa, C.I., Davis, M.F., Himmel, M.E., Adney, W.S., and Johnson, D.K. (2007). Cellulase digestibility of pretreated biomass is limited by cellulose accessibility. Biotechnol. Bioeng. *98*, 112–122.

Ji, S., Wang, S., Tan, Y., Chen, X., Schwarz, W., and Li, F. (2012). An untapped bacterial cellulolytic community enriched from coastal marine sediment under anaerobic and thermophilic conditions. Fems Microbiol. Lett. *335*, 39–46.

Jouany, J.-P. (2006). Optimizing rumen functions in the close-up transition period and early lactation to drive dry matter intake and energy balance in cows. Anim. Reprod. Sci. *96*, 250–264.

Kim, T.H., Lee, Y.Y., Sunwoo, C., and Kim, J.S. (2006). Pretreatment of corn stover by low-liquid ammonia recycle percolation process. Appl. Biochem. Biotechnol. *133*, 41–57.

Klein-Marcuschamer, D., Oleskowicz-Popiel, P., Simmons, B.A., and Blanch, H.W. (2012). The challenge of enzyme cost in the production of lignocellulosic biofuels. Biotechnol. Bioeng. *109*, 1083–1087.

Koeppel, A.F., and Wu, M. (2013). Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. Nucleic Acids Res. *41*, 5175–5188.

Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., and Schloss, P.D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl. Environ. Microbiol. *79*, 5112–5120.

Kumar, P.S., Brooker, M.R., Dowd, S.E., and Camerlengo, T. (2011). Target Region Selection Is a Critical Determinant of Community Fingerprints Generated by 16S Pyrosequencing. Plos One *6*.

Kunin, V., Engelbrektson, A., Ochman, H., and Hugenholtz, P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ. Microbiol. *12*, 118–123.

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat. Biotechnol. *31*, 814–821.

Laureano-Perez, L., Teymouri, F., Alizadeh, H., and Dale, B.E. (2005). Understanding factors that limit enzymatic hydrolysis of biomass. Appl. Biochem. Biotechnol. *124*, 1081–1099.

Leedle, J., Bryant, M., and Hespell, R. (1982). Diurnal-Variations in Bacterial Numbers and Fluid Parameters in Ruminal Contents of Animals Fed Low-Forage or High-Forage Diets. Appl. Environ. Microbiol. *44*, 402–412.

Li, M., Penner, G. b., Hernandez-Sanabria, E., Oba, M., and Guan, L. l. (2009). Effects of sampling location and time, and host animal on assessment of bacterial diversity and fermentation parameters in the bovine rumen. J. Appl. Microbiol. *107*, 1924–1934.

Lin, Z. (2013). Screw Extrusion Pretreatments to Enhance the Hydrolysis of Lignocellulosic Biomass. J. Microb. Biochem. Technol. *1*.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012a). Comparison of next-generation sequencing systems. J. Biomed. Biotechnol. *2012*, 251364.

Liu, W., Li, L., Khan, M.A., and Zhu, F. (2012b). Popular molecular markers in bacteria. Mol. Genet. Microbiol. Virol. *27*, 103–107.

Liu, Z., Lozupone, C., Hamady, M., Bushman, F.D., and Knight, R. (2007). Short pyrosequencing reads suffice for accurate microbial community analysis. Nucleic Acids Res. *35*.

Lozupone, C., and Knight, R. (2005). UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. Appl. Environ. Microbiol. *71*, 8228–8235.

Lozupone, C., Lladser, M.E., Knights, D., Stombaugh, J., and Knight, R. (2011). UniFrac: an effective distance metric for microbial community comparison. Isme J. *5*, 169–172.

Lü, Y., Li, N., Gong, D., Wang, X., and Cui, Z. (2012). The effect of temperature on the structure and function of a cellulose-degrading microbial community. Appl. Biochem. Biotechnol. *168*, 219–233.

Magnusson, L., Islam, R., Sparling, R., Levin, D., and Cicek, N. (2008). Direct hydrogen production from cellulosic waste materials with a single-step dark fermentation process. Int. J. Hydrog. Energy *33*, 5398–5403.

Magoč, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinforma. Oxf. Engl. *27*, 2957–2963.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2014). Swarm: robust and fast clustering method for amplicon-based studies. PeerJ *2*, e593.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12.

Martínez, M.E., Ranilla, M.J., Tejido, M.L., Saro, C., and Carro, M.D. (2010). Comparison of fermentation of diets of variable composition and microbial populations in the rumen of sheep and Rusitec fermenters. II. Protozoa population and diversity of bacterial communities. J. Dairy Sci. *93*, 3699–3712.

Masella, A.P., Bartram, A.K., Truszkowski, J.M., Brown, D.G., and Neufeld, J.D. (2012). PANDAseq: paired-end assembler for illumina sequences. BMC Bioinformatics *13*, 31.

McMurdie, P.J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PloS One *8*, e61217.

Menon, V., and Rao, M. (2012). Trends in bioconversion of lignocellulose: Biofuels, platform chemicals & biorefinery concept. Prog. Energy Combust. Sci. *38*, 522–550.

Michelland, R. j., Monteils, V., Zened, A., Combes, S., Cauquil, L., Gidenne, T., Hamelin, J., and Fortun-Lamothe, L. (2009a). Spatial and temporal variations of the bacterial community in the bovine digestive tract. J. Appl. Microbiol. *107*, 1642–1650.

Michelland, R.J., Dejean, S., Combes, S., Fortun-Lamothe, L., and Cauquil, L. (2009b). StatFingerprints: a friendly graphical interface program for processing and analysis of microbial fingerprint profiles. Mol. Ecol. Resour. *9*, 1359–1363.

Mikaelyan, A., Strassert, J.F.H., Tokuda, G., and Brune, A. (2014). The fibre-associated cellulolytic bacterial community in the hindgut of wood-feeding higher termites (Nasutitermes spp.). Environ. Microbiol. *16*, 2711–2722.

Mikaelyan, A., Dietrich, C., Köhler, T., Poulsen, M., Sillam-Dussès, D., and Brune, A. (2015a). Diet is the primary determinant of bacterial community structure in the guts of higher termites. Mol. Ecol. *24*, 5284–5295.

Mikaelyan, A., Köhler, T., Lampert, N., Rohland, J., Boga, H., Meuser, K., and Brune, A. (2015b). Classifying the bacterial gut microbiota of termites and cockroaches: A curated phylogenetic reference database (DictDb). Syst. Appl. Microbiol. *38*, 472–482.

Muggli, M.D., Puglisi, S.J., Ronen, R., and Boucher, C. (2015). Misassembly detection using paired-end sequence reads and optical mapping data. Bioinformatics *31*, i80–i88.

Munoz, R., Yarza, P., Ludwig, W., Euzéby, J., Amann, R., Schleifer, K.-H., Oliver Glöckner, F., and Rosselló-Móra, R. (2011). Release LTPs104 of the All-Species Living Tree. Syst. Appl. Microbiol. *34*, 169–170.

Muyzer, G. (1999). DGGE/TGGE a method for identifying genes from natural ecosystems. Curr. Opin. Microbiol. *2*, 317–322.

Mysara, M., Saeys, Y., Leys, N., Raes, J., and Monsieurs, P. (2015). CATCh, an ensemble classifier for chimera detection in 16S rRNA sequencing studies. Appl. Environ. Microbiol. *81*, 1573–1584.

Naz, S., and Fatima, A. (2013). Amplification of GC-rich DNA for high-throughput family-based genetic studies. Mol. Biotechnol. *53*, 345–350.

Nebel, M., Pfabel, C., Stock, A., Dunthorn, M., and Stoeck, T. (2011). Delimiting operational taxonomic units for assessing ciliate environmental diversity using small-subunit rRNA gene sequences. Environ. Microbiol. Rep. *3*, 154–158.

Nelson, M.C., Morrison, H.G., Benjamino, J., Grim, S.L., and Graf, J. (2014). Analysis, optimization and verification of Illumina-generated 16S rRNA gene amplicon surveys. PloS One *9*, e94249.

Nocker, A., Burr, M., and Camper, A.K. (2007). Genotypic microbial community profiling: A critical technical review. Microb. Ecol. *54*, 276–289.

Pedrós-Alió, C. (2012). The Rare Bacterial Biosphere. Annu. Rev. Mar. Sci. *4*, 449–466.

Peng, G., Zhu, W., Wang, H., Lue, Y., Wang, X., Zheng, D., and Cui, Z. (2010). Functional Characteristics and Diversity of a Novel Lignocelluloses Degrading Composite Microbial System with High Xylanase Activity. J. Microbiol. Biotechnol. *20*, 254–264.

Petri, R. m., Forster, R. j., Yang, W., McKinnon, J. j., and McAllister, T. a. (2012). Characterization of rumen bacterial diversity and fermentation parameters in concentrate fed cattle with and without forage. J. Appl. Microbiol. *112*, 1152–1162.

Pfennig, N., and Trüper, H.G. (1992). The Family Chromatiaceae. In The Prokaryotes, A. Balows, H.G. Trüper, M. Dworkin, W. Harder, and K.-H. Schleifer, eds. (Springer New York), pp. 3200–3221.

Pitta, D.W., Pinchak, W.E., Dowd, S.E., Osterstock, J., Gontcharova, V., Youn, E., Dorton, K., Yoon, I., Min, B.R., Fulford, J.D., et al. (2009). Rumen Bacterial Diversity Dynamics Associated with Changing from Bermudagrass Hay to Grazed Winter Wheat Diets. Microb. Ecol. *59*, 511–522.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. *41*, D590-596.

Quince, C., Lanzén, A., Curtis, T.P., Davenport, R.J., Hall, N., Head, I.M., Read, L.F., and Sloan, W.T. (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. Nat. Methods *6*, 639–641.

Quince, C., Lanzen, A., Davenport, R.J., and Turnbaugh, P.J. (2011). Removing noise from pyrosequenced amplicons. BMC Bioinformatics *12*, 38.

Reddy, A.P., Simmons, C.W., Claypool, J., Jabusch, L., Burd, H., Hadi, M.Z., Simmons, B.A., Singer, S.W., and VanderGheynst, J.S. (2012). Thermophilic enrichment of microbial communities in the presence of the ionic liquid 1-ethyl-3-methylimidazolium acetate. J. Appl. Microbiol. *113*, 1362–1370.

Roggenbuck, M., Sauer, C., Poulsen, M., Bertelsen, M.F., and Sørensen, S.J. (2014). The giraffe (Giraffa camelopardalis) rumen microbiome. FEMS Microbiol. Ecol.

Rollin, J.A., Zhu, Z., Sathitsuksanoh, N., and Zhang, Y.-H.P. (2011). Increasing cellulose accessibility is more important than removing lignin: a comparison of cellulose solvent-based lignocellulose fractionation and soaking in aqueous ammonia. Biotechnol. Bioeng. *108*, 22–30.

Schirmer, M., Ijaz, U.Z., D'Amore, R., Hall, N., Sloan, W.T., and Quince, C. (2015). Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res.

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., et al. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. Appl. Environ. Microbiol. *75*, 7537–7541.

Schloss, P.D., Gevers, D., and Westcott, S.L. (2011). Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PloS One *6*, e27310.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W., Consortium, F.B., List, F.B.C.A., Bolchacova, E., et al. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc. Natl. Acad. Sci. *109*, 6241–6246.

Schwieger, F., and Tebbe, C.C. (1998). A new approach to utilize PCR-single-strand-conformation polymorphism for 16s rRNA gene-based microbial community analysis. Appl. Environ. Microbiol. *64*, 4870–4876.

Shin, S.C., Ahn, D.H., Kim, S.J., Lee, H., Oh, T.-J., Lee, J.E., and Park, H. (2013). Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes. PLoS ONE *8*, e68824.

Simmons, C. w., Reddy, A. p., Simmons, B. a., Singer, S. w., and VanderGheynst, J. s. (2014). Effect of inoculum source on the enrichment of microbial communities on two lignocellulosic bioenergy crops under thermophilic and high-solids conditions. J. Appl. Microbiol. *117*, 1025–1034.

Spor, A., Koren, O., and Ley, R. (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. Nat. Rev. Microbiol. *9*, 279–290.

Sun, Y., and Cheng, J. (2002). Hydrolysis of lignocellulosic materials for ethanol production: a review. Bioresour. Technol. *83*, 1–11.

Sundset, M.A., Præsteng, K.E., Cann, I.K.O., Mathiesen, S.D., and Mackie, R.I. (2007). Novel Rumen Bacterial Diversity in Two Geographically Separated Sub-Species of Reindeer. Microb. Ecol. *54*, 424–438.

Sundset, M.A., Edwards, J.E., Cheng, Y.F., Senosiain, R.S., Fraile, M.N., Northwood, K.S., Praesteng, K.E., Glad, T., Mathiesen, S.D., and Wright, A.-D.G. (2009). Rumen microbial diversity in Svalbard reindeer, with particular emphasis on methanogenic archaea. FEMS Microbiol. Ecol. *70*, 553–562.

Tajima, K., Aminov, R.I., Nagamine, T., Ogata, K., Nakamura, M., Matsui, H., and Benno, Y. (1999). Rumen bacterial diversity as determined by sequence analysis of 16S rDNA libraries. FEMS Microbiol. Ecol. *29*, 159–169.

Todaka, N., Inoue, T., Saita, K., Ohkuma, M., Nalepa, C.A., Lenz, M., Kudo, T., and Moriya, S. (2010). Phylogenetic Analysis of Cellulolytic Enzyme Genes from Representative Lineages of Termites and a Related Cockroach. PLoS ONE *5*, e8636.

Tokuda, G., and Watanabe, H. (2007). Hidden cellulases in termites: revision of an old hypothesis. Biol. Lett. *3*, 336–339.

Tokuda, G., Lo, N., Watanabe, H., Arakawa, G., Matsumoto, T., and Noda, H. (2004). Major alteration of the expression site of endogenous cellulases in members of an apical termite lineage. Mol. Ecol. *13*, 3219–3228.

Torella, J.P., Ford, T.J., Kim, S.N., Chen, A.M., Way, J.C., and Silver, P.A. (2013). Tailored fatty acid synthesis via dynamic control of fatty acid elongation. Proc. Natl. Acad. Sci. *110*, 11290–11295.

Vassilev, S.V., Baxter, D., Andersen, L.K., Vassileva, C.G., and Morgan, T.J. (2012). An overview of the organic and inorganic phase composition of biomass. Fuel *94*, 1–33.

Wallace, R.J., Rooke, J.A., McKain, N., Duthie, C.-A., Hyslop, J.J., Ross, D.W., Waterhouse, A., Watson, M., and Roehe, R. (2015). The rumen microbial metagenome associated with high methane production in cattle. BMC Genomics *16*, 839.

Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. *73*, 5261–5267.

Warnecke, F., Luginbuehl, P., Ivanova, N., Ghassemian, M., Richardson, T.H., Stege, J.T., Cayouette, M., McHardy, A.C., Djordjevic, G., Aboushadi, N., et al. (2007). Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature *450*, 560-U17.

Weimer, P.J., Stevenson, D.M., Mantovani, H.C., and Man, S.L.C. (2010). Host specificity of the ruminal bacterial community in the dairy cow following near-total exchange of ruminal contents1. J. Dairy Sci. *93*, 5902–5912.

Wen, B., Yuan, X., Li, Q.X., Liu, J., Ren, J., Wang, X., and Cui, Z. (2015). Comparison and evaluation of concurrent saccharification and anaerobic digestion of Napier grass after pretreatment by three microbial consortia. Bioresour. Technol. *175*, 102–111.

Werner, J.J., Zhou, D., Caporaso, J.G., Knight, R., and Angenent, L.T. (2012). Comparison of Illumina paired-end and single-direction sequencing for microbial 16S rRNA gene amplicon surveys. ISME J. *6*, 1273–1276.

Winsley, T., van Dorst, J.M., Brown, M.V., and Ferrari, B.C. (2012). Capturing Greater 16S rRNA Gene Sequence Diversity within the Domain Bacteria. Appl. Environ. Microbiol. *78*, 5938–5941.

Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. U. S. A. *74*, 5088–5090.

Wongwilaiwalin, S., Rattanachomsri, U., Laothanachareon, T., Eurwilaichitr, L., Igarashi, Y., and Champreda, V. (2010). Analysis of a thermophilic lignocellulose degrading microbial consortium and multi-species lignocellulolytic enzyme system. Enzyme Microb. Technol. *47*, 283–290.

Xu, M., Schnorr, J., Keibler, B., and Simon, H.M. (2012). Comparative Analysis of 16S rRNA and amoA Genes from Archaea Selected with Organic and Inorganic Amendments in Enrichment Culture. Appl. Environ. Microbiol. *78*, 2137–2146.

Xue, Z., Zhang, W., Wang, L., Hou, R., Zhang, M., Fei, L., Zhang, X., Huang, H., Bridgewater, L.C., Jiang, Y., et al. (2015). The Bamboo-Eating Giant Panda Harbors a Carnivore-Like Gut Microbiota, with Excessive Seasonal Variations. mBio *6*, e00022-15.

Yu, Y., Lee, C., Kim, J., and Hwang, S. (2005). Group-specific primer and probe sets to detect methanogenic communities using quantitative real-time polymerase chain reaction. Biotechnol. Bioeng. *89*, 670–679.

Zhao, H., Yu, H., Yuan, X., Piao, R., Li, H., Wang, X., and Cui, Z. (2014a). Degradation of Lignocelluloses in Rice Straw by BMC-9, a Composite Microbial System. J. Microbiol. Biotechnol. *24*, 585–591.

Zhao, X., Wang, L., Lu, X., and Zhang, S. (2014b). Pretreatment of corn stover with diluted acetic acid for enhancement of acidogenic fermentation. Bioresour. Technol. *158*, 12–18.

Zuckerkandl, E., and Pauling, L. (1965). Molecules as documents of evolutionary history. J. Theor. Biol. *8*, 357–366.