THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (Toulouse INP)

Discipline ou spécialité :

Pathologie, Toxicologie, Génétique et Nutrition

Présentée et soutenue par :

M. MALO LE BOULCH

le vendredi 20 décembre 2019

Titre :

Taxonomie et inférence fonctionnelle des procaryotes: développement de MACADAM, une base de données devoies métaboliques associées à une taxonomie

---

Ecole doctorale :

Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingénieries (SEVAB)

Unité de recherche :

Génétique, Physiologie et Systèmes d'Elevage (GENPHYSE)

Directeur(s) de Thèse :

MME GERALDINE PASCAL

MME SYLVIE COMBES

Rapporteurs :

M. DIDIER DEBROAS, UNIVERSITE CLERMONT AUVERGNE

M. PIERRE PEYRET, UNIVERSITE CLERMONT AUVERGNE

Membre(s) du jury :

Mme JULIETTE RIQUET, INRA TOULOUSE, Président

Mme GERALDINE PASCAL, INRA TOULOUSE, Membre

Mme MARION LECLERC, INRA JOUY EN JOSAS, Membre

Mme SYLVIE COMBES, INRA TOULOUSE, Membre

"I travel not to go anywhere, but to go. I travel for travel's sake. The great affair is to move."
par Robert Louis Stevenson

Remerciements

Je tiens en tout premier lieu à remercier les deux personnes ayant accepté d'être rapporteurs de ma thèse : le professeur Pierre Peyret et le professeur Didier Debroas. J'adresse également mes remerciements à Marion Leclerc de faire partie de mon jury dans l'évaluation de travaux effectués lors de cette thèse.

Je tiens à remercier vivement mes deux directrices de thèse, Géraldine Pascal et Sylvie Combes. Merci de m'avoir permis de réaliser cette thèse avec vous et d'avoir été d'un soutien, scientifique ou personnel, sans faille pendant ces trois ans. Ceux-ci ont été remplis de nouvelles expériences et de nouveaux savoirs et forment une aventure de vie que je ne suis pas près de mettre de coté que je ne suis pas près d'oublier. Encore une fois, merci.

Je voudrais ensuite remercier les membres de mon comité de thèse : Stéphane Uroz, Nicolas Derome et Jordi Estelle-Fabrellas. Je remercie chaleureusement aussi Patrice Dehais, pour son aide sur MACADAM et sa bonne humeur perpétuelle dans le cadre de nos travaux. J'en profite pour remercier tout le personnel de Genotoul et en particulier Marie-Stéphane Trotard et Didier Laborie. Je tiens également à remercier Cédric Cabau pour, en plus de sa présence dans mon comité de thèse, son aide pour le site web de MACADAM.

Je tiens ensuite à remercier le directeur de GenPhyse lors de mon arrivée et son successeur : Xavier Fernandez et Juliette Riquet. Par la même occasion, j'en profite pour remercier les différents financeurs de ce projet, à savoir le département PHASE et la région Occitanie sans qui ces travaux n'auraient pu avoir lieu. Je tiens à remercier également DARESE et l'université Paul Sabatier pour le financement de mon séjour de trois mois en Italie dans le cadre du parcours doctoral de l'École internationale de recherche d'Agreenium.

Je tiens ensuite à remercier tout le couloir E2 composé des équipes NED et SYSED. Merci à Annabelle, Olivier, Caroline, Béatrice, Thierry, Laurence et Muriel. Merci pour cette ambiance exceptionnelle et ces conversations enrichissantes. Ensuite, je souhaite remercier Laurent et Davi pour les nombreuses conversations diverses et variées et notamment sport tout le long de ma thèse. Je souhaite remercier ensuite le trio de choc du bureau E208 : Carole, Céline et Manon dont la porte a toujours été ouverte. Je souhaite ensuite remercier Martin pour la relecture de la partie métabolique de cette thèse et ses apports loin d'être négligeables dans celle-ci. Je remercie ensuite tous les gens qui ont partagé mon bureau (ou pas) au cours de ces trois ans et ont subi mes trop grandes discussions (dans l'ordre d'ancienneté d'arrivée) : Christelle (bien que nous n'ayons pas été ensemble elle fait un peu partie du

bureau des doctorants), Héloïse et Clémentine (#Nakache4ever), Yayu pour son calme, Charlotte pour sa bonne humeur permanente, Cécile et Eloïse. Je souhaite enfin remercier des gens qui ont été de passage dans ce couloir mais qui sont partis vers d'autres horizons : Marie-Lea, Noémie, Nathalie, Iván et tous les stagiaires. J'en profite également pour remercier Louise et Émilie du couloir C pour la bonne ambiance en séminaire EIR-A ainsi que Manuela, Florence, Nancy et Evelyne.

Il fut une période de travaux où un changement de bureau a été nécessaire et je remercie donc le directeur de l'unité Sylvain Jasson pour son accueil et Fulya pour m'avoir accueilli dans son bureau. J'aimerais remercier les personnes de MIAT avec qui j'ai pu partager des pauses thé très intéressantes scientifiquement. Je tiens également à remercier Manon (j'attends toujours notre combat à l'aéroport Blagnac). Je remercie Lise (c'était vachement bien Marseille et Athènes hein ?). Je remercie ensuite particulièrement Damien, Émilien et Floréal pour les nombreux moments de sensibilisation que nous avons eu.

Je tiens à remercier le professeur Nicola Segata de m'avoir accueilli dans son équipe le temps de trois mois afin d'accentuer mes connaissances en bioinformatique. Merci aux membres de son équipe pour avoir facilité mon séjour notamment à Adrian, Francesco et Francesco, Serena, Kun, Moreno, Paolo et Nicolai.

Voici enfin les remerciements plus personnels. Merci aux gens m'ayant soutenu à distance tous les soirs en jouant sûrement trop et en partageant de nombreuses sorties lors de vacances : Maël, Aymeric, Florent, Gabriel, Nico, J-M, Yohann, Charles. Merci à Dominique pour cette première soirée sur Toulouse. Merci à tous les amis qui sont sur Rennes : Pierre, Damien R, Marianne, Nico T, Damien R, Cindy, Fabien, Gaëtan, David H, David J, Elisa, Ismaël, Christelle, et à ceux qui ne sont pas sur Rennes : Julie et Gaelle. J'en oublie sûrement dans ce cas excusez-moi. Merci au chan #master_msb pour leur soutien techniques et les discussions enrichissantes particulièrement à Sylvain et Nico M.

Les derniers remerciements de cette thèse iront bien sûr aux membres de ma famille, ma mère, mon père, ma soeur, Vincent et tous ceux qui ne sont malheureusement plus présents. Merci à vous pour votre soutien pendant ces trois ans. Enfin merci à Victoria de m'avoir supporté, remonté le moral et re-supporté pendant ces trois ans. Sans toi rien de tout ça n'aurait pu se réaliser.

Merci à toi enfin personne que j'ai sûrement oublié dans mes remerciements.

Table des figures

Liste des tableaux

Liste des abréviations, des sigles et des symboles

AAI Average Amino acid Identity (moyenne d'identité des acides aminés)

ADN Acide DésoxyriboNucléique

ANI Average Nucleotide sequence Identity (Identité moyenne de la séquence nucléotidique)

API Application Programming Interface

ARN Acide ribonucléique

ASV Amplicon Sequence Variant

BRENDA BRaunschweig ENzyme DAtabase

DACU Dernier Ancêtre Commun Universel

DIAMOND Double Index AlignMent Of Next-generation sequencing Data

DNA Data Bank of Japan DDBJ

EBI European Bioinformatics Institute

EC number Nomenclature EC

EMBL European Molecular Biology Laboratory

GTDB Genome Taxonomy DataBase

HMP Human Microbiome Project

HUMAnN2 The HMP Unified Metabolic Analysis Network 2

ICN International Code of Nomenclature for algae, fungi, and plants

ICNP International Code of Nomenclature of Prokaryotes

ICSP International Committee for Systematic of Prokaryotes

ICZN International Commission on Zoological Nomenclature

IJSEM International Journal of Systematic and Evolutionary Microbiology

IMG/M The Integrated Microbial Genomes & Microbiomes

InChI IUPAC International Chemical Identifier

INSDC International Nucleotide Sequence Database Collaboration

IUPAC International Union of Pure and Applied Chemistry

KEGG The Kyoto Encyclopedia of Genes and Genomes

KO KEGG Orthologs

LPSN The List of Prokaryotic Names with Standing in Nomenclature

LTP Living Tree Project

LTP The All-Species Living Tree Project

MAG Metagenome-Assembled Genomes

MGS MetaGenomic Species

MLSA Multi-Locus Sequence Analysis

MLST Multi-Locus Sequence Typing

MSA Multiple Sequence Alignment

NCBI National Center for Biotechnology Information

OTL Open Tree of Life

OTT OpenTree Taxonomy

OTU Unité Taxonomique Opérationnelle

PAPRICA PAthway PRediction by phylogenetIC plAcement

PATRIC Pathosystems Resource Integration Center

PFS Pathway Frequency Score

PGDB Pathway/Genome DataBase

PICRUSt Phylogenetic Investigation of Communities by Reconstruction of Unobserved States

PIR Protein Information Ressource

PS Pathway Score

RDP The Ribosomal Database Project

RPK Lectures par kilobase (Reads per Kilobase)

rrnDB Ribosomal RNA Database

SGB Species-level Genome Bins

SIB Swiss Institute of Bioinformatics

SRA Sequence Read Archive

UniProt UNIversal PROTein reference

UniRef UniProt Reference Clusters

Productions scientifiques

Publications scientifiques

— M. Le Boulch, P. Déhais, S. Combes et G. Pascal (1er jan. 2019). « The MACADAM database : a MetAboliC pAthways DAtabase for Microbial taxonomic groups for mining potential metabolic capacities of archaeal and bacterial taxonomic groups ». In : *Database* 2019. doi : 10.1093/database/baz049

— T. Read, L. Fortun-Lamothe, G. Pascal, M. Le Boulch, L. Cauquil, B. Gabinaud, C. Bannelier, E. Balmisse, N. Destombes, O. Bouchez, T. Gidenne et S. Combes (2019). « Diversity and Co-occurrence Pattern Analysis of Cecal Microbiota Establishment at the Onset of Solid Feeding in Young Rabbits ». In : *Frontiers in Microbiology* 10. issn : 1664-302X. doi : 10.3389/fmicb.2019.00973

— H. Falentin, L. Auer, M. Mariadassou, G. Pascal, O. Rué, E. DUGAT-BONY, C. Delbes, A. Nicolas, E. RIFA, S. MONDY, M. Le Boulch, L. Cauquil, G. Hernandez Raquet, S. Terrat et A.-L. Abraham (2019). « Guide pratique à destination des biologistes, bioinformaticiens et statisticiens qui souhaitent s'initier aux analyses métabarcoding ». In : *Cahiers des Techniques de l'INRA* 2019.97, p. 1-23

Communications orales

— M. Le Boulch, S. Combes et G. Pascal (16 mai 2017). « Functional inference of complex bacterial communities from marker genes derived from high-throughput sequencing ». Présentation orale. Présentation orale. NEM 2017. Saint Pée-sur-Nivelle

— M. Le Boulch, S. Combes et G. Pascal (16 mai 2018). « Difficultés de l'inférence fonctionnelle à partir du 16S et présentation de MACADAM ». Présentation orale. Présentation orale. NEM 2018. Narbonne

— M. Le Boulch, P. Dehais, S. Combes et G. Pascal (7 avr. 2018a). « MACADAM a user-friendly MetAboliC pAthway DAtabase for complex Microbial community function analysis ». Présentation orale. Présentation orale. JOBIM 2018. Marseille

Poster

— M. Le Boulch, P. Dehais, S. Combes et G. Pascal (11 sept. 2018b). « MACADAM : a user-friendly MetAboliC pAthway DAtabase for complex Microbial community function analysis ». Poster. Poster. ECCB 2018. Athènes

Site internet

— M. Le Boulch, P. Déhais, S. Combes et G. Pascal (2018c). *MACADAMExplore*. url : http://macadam.toulouse.inra.fr/

Table des matières

Introduction générale

Les micro-organismes ont été formellement observés la première fois par Antonie van Leeuwenhoek en 1676 avec un microscope de sa création (Leeuwenhoek 1677). Les micro-organismes sont composés, entre autres, des bactéries et des archées, qui forment le groupe d'organismes le plus ubiquitaire de notre planète : les procaryotes (Stanier et Niel 1962). Ils ont colonisé tous les biomes, le sol (Fierer, Bradford et Jackson 2007), l'eau (Louca, Parfrey et Doebeli 2016) et l'atmosphère (Bauer et al. 2003), mais également les milieux les plus extrêmes en termes de pression, de température et/ou de radiations (Stetter 1996). Avec l'apparition des organismes eucaryotes multicellulaires, les communautés procaryotiques ont rapidement colonisé ces hôtes et développé des stratégies conduisant à la création de mutualisme ou de symbiose.

Les procaryotes présentent une grande plasticité génomique et une extrême diversité métabolique en lien avec leur omniprésence dans ces environnements (Oren 2009). L'homme a appris très tôt à se servir de ces capacités métaboliques à son avantage (Hanniffy et al. 2009) et à chercher à mieux caractériser ces organismes dans un objectif de maîtrise de leur fonctionnement. Pour rendre intelligible la vaste diversité du monde procaryotique et comprendre les relations entre les individus, identifier, nommer et classer sont alors des étapes incontournables. On fait alors appel à la taxonomie. La taxonomie est une science en perpétuelle évolution, qui a connu et connaît encore de profonds bouleversements au rythme des progrès de la biologie, passant de critères phénotypiques aux avancées de la biologie moléculaire et des techniques de séquençage à haut débit (Kämpfer et Glaeser 2012). Il est désormais possible de découvrir de nouvelles espèces sans observations phénotypiques, sans culture microbiologique, ni connaissance préalable de leurs caractéristiques métaboliques (Pasolli et al. 2019). Ces découvertes conduisent à des remaniements de la classification sans pour autant résoudre la définition de l'espèce qui reste une difficulté majeure chez les procaryotes (Rosselló-Móra et Amann 2015). Dans ce contexte d'afflux de données métagénomiques conduisant à l'identification de nouvelles espèces, il appa-

raît fondamental de questionner les résultats d'affiliation taxonomique issus de chaînes de traitements automatisés.

La volonté d'identifier et de classer les organismes est étroitement associée à la nécessité de comprendre - voire de maîtriser - le fonctionnement des procaryotes au sein d'une communauté. Les approches par séquençage de l'ADN à l'échelle de l'individu ou à l'échelle massive des communautés (métagénomique) donnent accès au potentiel fonctionnel des individus et communautés procaryotes (Bengtsson-Palme 2018). Plusieurs méthodologies existent afin d'inférer les fonctions d'une communauté procaryote dépendant ou non de l'affiliation taxonomique préalable de ces membres (Franzosa et al. 2018 ; Asshauer et al. 2015). Mais cette détermination du potentiel fonctionnel par inférence fonctionnelle repose à la fois sur la disponibilité de génomes ayant une qualité de séquençage et d'assemblage suffisante, et sur l'utilisation de banque de données répertoriant les voies métaboliques et leurs fonctions. L'une des difficultés actuelles dans la détermination du potentiel fonctionnel d'un organisme est de pouvoir disposer de données fiables et qui évoluent en fonction des derniers développements technologiques et de l'avancée des connaissances.

La première partie de ce manuscrit est consacrée à une synthèse bibliographique portant sur la définition d'un procaryote, le concept d'espèces, et les spécificités de la taxonomie procaryote. Ensuite, nous décrirons les méthodes d'identification des organismes dans une communauté bactérienne. Enfin, nous aborderons les grand principes du métabolisme procaryote et de l'inférence fonctionnelle. Dans une deuxième partie, nous présenterons les résultats d'affiliation taxonomique d'espèces métagénomiques dans les genres *Blautia* et *Ruminococcus*, en mettant en oeuvre des techniques génétiques et phylogénétiques. Ces travaux ont été effectués au cours d'un séjour dans l'équipe du Professeur Nicola Segata de l'université de Trente en Italie, dans le cadre de la labellisation Agreenium de cette thèse. Ensuite, nous présenterons le développement de la base de données MACADAM, tâche majeure de ce travail de thèse. MACADAM est une base de données liant taxonomie et informations fonctionnelles, afin d'inférer le potentiel fonctionnel de taxons. Cette description sera suivie de deux exemples d'utilisation de MACADAM. Enfin, nous comparerons l'inférence de voie métabolique à partir de lectures métagénomiques et à partir de données taxonomiques. L'ensemble des résultats obtenus, ainsi que les avantages et limites de MACADAM, seront discutés.

Chapitre A

État de l'art

A.1 Caractérisation des espèces procaryotes : intérêts et limites

A.1.1 Définition et importance

Les micro-organismes (du grec *mikrós* : petit) unicellulaires, également appelés microbes, ont été les premiers organismes à se développer sur terre, il y a 3,5 milliards d'années (Cavalier-Smith 2006). Le phylum des *Chloroflexi* (anciennement appelé *Chlorobacteria*) serait le phylum bactérien le plus proche (Valas et Bourne 2009) du Dernier Ancêtre Commun Universel (DACU, Ouzounis et Kyrpides 1996) pouvant être défini comme le dernier organisme avant la bifurcation qui a donné naissance aux domaines des *Bacteria* (ou *Eubacteria*) et des *Archaea* (Cornish-Bowden et Cárdenas 2017). C'est dans ce dernier qu'aura lieu l'émergence du domaine des *Eukaryota* (Dacks et al. 2016). Cette classification à trois domaines (Figure A.I) a été proposée en 1990 par Carl Woese (Woese et Fox 1977 ; Woese, Kandler et Wheelis 1990) grâce aux avancées des méthodes de biologie moléculaire et à l'étude de l'ARN ribosomique 16S, ARN constitutif de la petite sous unité du ribosome.

Les micro-organismes sont présents dans les trois domaines de cette classification. Les domaines des *Bacteria* et des *Archaea* sont exclusivement composés de micro-organismes unicellulaires. Ils sont également présents dans le domaine des *Eukaryota* plus précisément dans le royaume des *Plantae* tel que les algues vertes ou dans le royaume des *Fungi* tel que les levures du genre *Saccharomyces*. Les autres micro-organismes unicellulaires eucaryotes sont appelés protistes. Selon l'une des définitions proposées par O'Malley, Simpson et Roger (2013), les protistes se définissent comme un groupe pa-

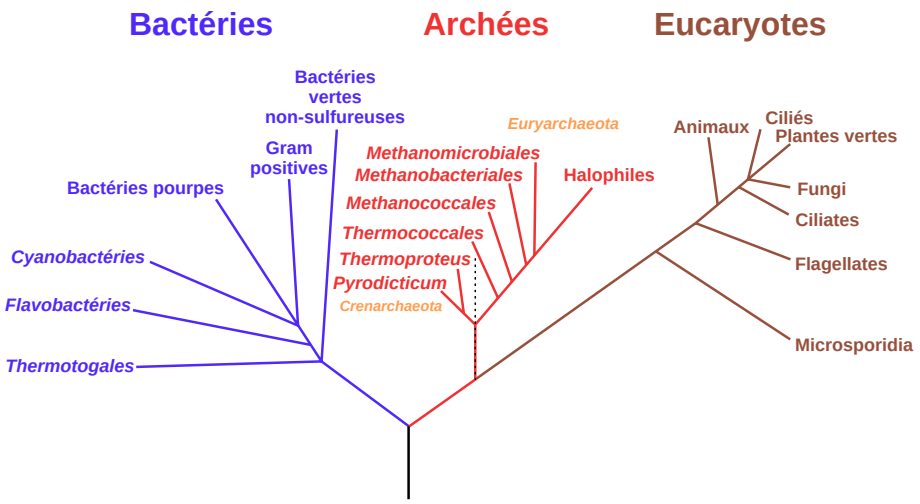


FIGURE A.I – Arbre du vivant en trois domaines.

D'après WOESE, KANDLER et WHEELIS (1990)

raphylétique composé des organismes eucaryotes n'étant jamais sous forme multicellulaire et n'appartenant pas aux royaumes *Fungi* et *Plantae*. Dans ce travail, nous ne traiterons que des micro-organismes unicellulaires appartenant aux domaines des *Bacteria* et des *Archaea*. Cet ensemble d'organismes est appelé procaryotes.

Selon STANIER et NIEL (1962), trois critères majeurs permettent de définir la cellule procaryote : (i) l'absence de membrane interne qui compartimente la cellule en séparant le matériel nucléaire de la machinerie enzymatique pour la respiration et/ou la photosynthèse, (ii) une division cellulaire qui s'effectue par scissiparité et non par mitose, et (iii) la présence de peptidoglycane dans la membrane cellulaire (pseudopeptidoglycane chez les archées). Ces critères ont été définis alors que procaryotes était un synonyme de bactérie. Mais à la suite de la séparation des domaines des bactéries et des archées, les procaryotes sont devenus un groupe paraphylétique défini dans le livre *The Prokaryotes* (ROSENBERG et al. 2013) comme étant un groupe constitué d'organismes appartenant à deux des trois domaines du vivant : celui des *Bacteria* et des *Archaea*. Les archées (anciennement appelées archaebactéries) étaient connues comme des bactéries extrêmophiles peuplant les grands fonds marins ou les lacs acides ou salés. Des études ont démontré que celles-ci n'appartenaient pas au domaine des bactéries mais formaient un domaine à part entière (WOESE, KANDLER et WHEELIS 1990) et dont les organismes sont présents dans tous les milieux (ROBERTSON et al. 2005 ; CHABAN, NG et JARRELL 2006) au même titre que les bactérie mais en plus

faible nombre. Les archées présentent une morphologie proche des bactéries tandis qu'une partie de leur machinerie cellulaire s'apparente à celle de la cellule eucaryote (Tableau A.1). Elles sont les seules formes de vie à posséder les voies métaboliques de la méthanogènese (Cavicchioli 2011).

Traits	**Bactéries**	**Archées**	**Eucaryotes**
Liaison carbone des lipides	Esther	Ether	Esther
Organisation des têtes phosphates des lipides	Glycerol-3-phosphate	Glycerol-1-phosphate	Glycerol-3-phosphate
Métabolisme	Bactérien	Similaire aux Bactéries	Eucaryote
Mécanisme de transcription de l'ADN	Bactérien	Similaire aux Eucaryotes	Eucaryote
Traduction des facteurs d'élongation	Bactérien	Similaire aux Eucaryotes	Eucaryote
Taille de l'ARN ribosomique de la petite sous-unité ribosomique	16S	16S	18S
Noyau	non	non	oui
Compartiment cellulaire	non	non	oui
Methanogénèse	non	oui	non
Photosynthèse chlorophyllienne	présente	absente	présente

Tableau A.1 – Quelques traits caractéristiques des domaines *Bacteria*, *Archaea* et *Eukaryota*.

D'après Lavergne (2014) et Cavicchioli (2011)

Les procaryotes ont investi tous les milieux : sol, eau, glace, racines, systèmes digestifs des animaux, entre autres, et catalysent des réactions uniques et indispensables aux cycles biogéochimiques de la biosphère ainsi que la production des composants essentiels de l'atmosphère terrestre. Leur nombre est compris entre 4 et 6×10^{30} cellules. L'ensemble des procaryotes constitue un réservoir contenant une quantité de carbone entre 60 à 100% de celle des plantes, et dix fois plus important en ce qui concerne l'azote et le phosphore

(5–130 $\times$ 10^{15}g et 9–14 $\times$ 10^{15}g respectivement, Whitman, Coleman et Wiebe 1998). Ils sont également à la base de tous les processus biogéochimiques tels que la production des gaz de l'atmosphère terrestre (Whitman 2009). Cette omniprésence est liée à l'extrême diversité des procaryotes.

A.1.2 Les concepts de l'espèce bactérienne

L'espèce est le taxon de base de la classification du vivant qui est sujet à de nombreuses discussions quant à son concept ou même son existence biologique (Ereshefsky 1992 ; Wilson 1999). Le concept de l'espèce le plus accepté par la communauté de biologiste est celui proposé par Mayr (1942) : une espèce biologique est une population ou un ensemble de populations dont les individus peuvent se reproduire entre eux et engendrer une descendance viable et féconde dans les conditions naturelles. Le concept d'espèce chez les bactéries [1] se complexifie, car ceux-ci se reproduisent de manière asexuelle par scissiparité. La scissiparité permet le transfert vertical de matériel génétique, ainsi d'un même organisme, deux clones sont obtenus. Toutefois, les transferts horizontaux entre bactéries d'espèces différentes constituent la principale source de transfert de matériel génétique (Treangen et Rocha 2011). Plusieurs concepts de l'espèce bactérienne sont proposés dans la littérature : (i) le concept d'espèce recombinante, (ii) le concept d'espèce écologique, (iii) le concept d'espèce unitaire sans méthode, (iv) le concept d'espèce phylogénétique et (v) le concept nominaliste d'espèce.

Le concept d'espèce recombinante Cette approche suggère que la délimitation d'une espèce bactérienne est complexe du fait de la constante recombinaison de son génome (Dykhuizen et Green 1991) par transferts horizontaux de matériel génétique selon trois mécanismes (Figure A.II, Paul 1999 ; Thomas et Nielsen 2005) : (i) la conjugaison, via contact entre cellules, (ii) la transformation, via l'incorporation de matériel génétique exogène provenant de son environnement, et (iii) la transduction, via le transfert du

1. Les *Archaea* ont été considérées comme des bactéries extrêmophiles appartenant au domaine des bactéries. Ce n'est qu'à partir de 1990 que celles-ci ont été isolées dans leur propre domaine (Woese, Kandler et Wheelis 1990). De nombreux travaux effectués avant et au cours de cette période parlent donc d'espèce bactérienne pour inclure les *Archaea* et les bactéries (l'ensemble formant les procaryotes). Notre travail se basant majoritairement sur les bactéries, nous n'utiliserons pas le terme d'espèce procaryotiques mais celui d'espèce bactérienne car, bien que de nombreux concepts tel que celui de l'espèce soient transposables aux *Archaea*, celles-ci n'ont pas été étudiées en détail dans notre travail.

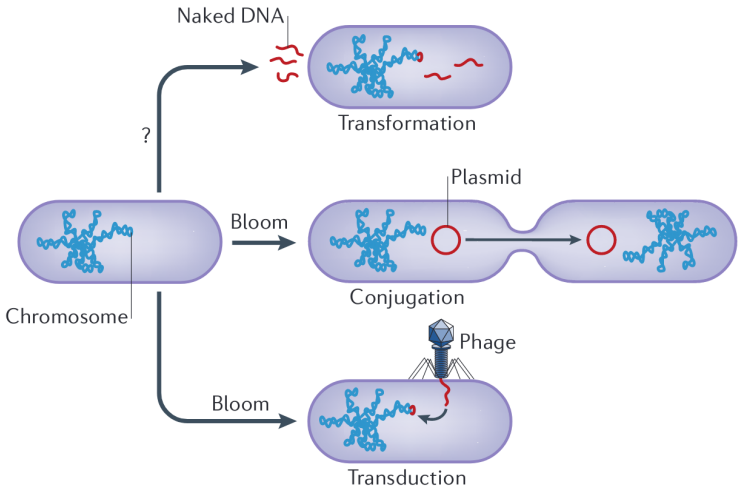


FIGURE A.II – Mécanismes de transferts horizontaux de gènes chez les bactéries.

Source: STECHER, MAIER et HARDT (2013)

matériel génétique par un bactériophage entre cellules.

Les transferts horizontaux sont facilités entre organismes présentant des génomes peu divergents, mais ils sont aussi possibles entre des organismes de niveau taxonomique éloigné (même au niveau des domaines GOGARTEN et TOWNSEND 2005). Ces échanges ont pour conséquence une diversification des populations bactérienne, mais rendent difficile - voire impossible - le concept d'espèce bactérienne. Il est toutefois possible d'identifier et de prévoir cette diversification des lignées (LAWRENCE 2002). Le concept d'espèce recombinante a été mis à mal par la mise en évidence de certains taxons où les recombinaisons (c.-à-d. les transferts horizontaux) sont rares (DYKHUIZEN et BARANTON 2001) et les transferts de matériel génétique seraient majoritairement verticaux. Ce concept serait donc limité aux taxons apparentés ou aux organismes d'espèces différentes vivant dans les mêmes environnements (BEIKO, HARLOW et RAGAN 2005).

Le concept d'espèce écologique Le concept d'espèce écologique a été proposé par COHAN (2002) : « Une espèce dans le monde bactérien peut être comprise comme une lignée évolutive liée par une sélection périodique spécifique à l'écotype », où l'écotype est défini comme « Un ensemble de souches vivant dans la même niche écologique de tel sorte que si un mutant mieux adapté apparaît, alors il marginalise jusqu'à l'extinction les souches de la niche écologique dont il est issu ». Selon ce concept, les espèces bacté-

riennes sont adaptées à leur environnement, et cet environnement applique des vagues de forces de sélection permettant d'épurer la divergence des organismes composant cette espèce. Les traits phénotypiques les plus adaptés sont conservés, tandis que ceux moins adaptés disparaissent. Selon cette approche, les espèces que nous connaissons à l'heure actuelle sont à rapprocher d'un genre plus que d'une espèce bactérienne. Ce concept d'espèce écologique est invalidé si les souches composant l'écotype tendent plus à transférer leur matériel génétique par transfert horizontal que par transfert vertical. Ces transferts horizontaux pourraient permettre l'adaptation des écotypes à ces vagues de pressions de sélection (Fraser, Hanage et Spratt 2007).

Le concept d'espèce unitaire sans méthode Le concept d'espèce unitaire sans méthode est défini comme un concept où les espèces sont des lignées de métapopulations (Queiroz 2005 ; Achtman et Wagner 2008). Une métapopulation est un ensemble de populations réparties dans l'espace ou dans le temps, entre lesquelles il existe des échanges (d'individus et/ou de matériel génétique). Une lignée de métapopulation est le développement de cette métapopulation au cours du temps. La seule contrainte de ce concept d'espèce est donc que la lignée de métapopulation évolue séparément des autres lignées de métapopulations. Ce concept est très large au niveau des individus qu'il englobe et ne semble pas permettre de classifier les espèces qu'il définit dans un système universel.

Le concept d'espèce phylogénétique Ce concept se base sur l'ADN afin de déterminer les relations phylogénétiques entre les différents organismes. Ici les espèces sont définies comme des clades, c'est à dire des groupes monophylétiques (groupe où tous les organismes descendent d'un même ancêtre commun Rosselló-Mora et Amann 2001). Cette approche utilise le matériel génétique présent dans les organismes pour définir des relations phylogénétiques. Le matériel génétique et les techniques utilisés peuvent être : (i) le gène de l'ARNr 16S (Woese et Fox 1977 ; Woese 1987 ; Stackebrandt et al. 2002), (ii) l'hybridation ADN/ADN de l'intégralité du génome de l'organisme (Marmur et Doty 1962 ; Grimont 1988), (iii) la séquence du génome afin de calculer le score ANI (Average Nucleotide sequence Identity) (Konstantinidis et Tiedje 2005a), (iv) les gènes composant le pangénome (Rosselló-Mora et Amann 2001) de l'espèce. Celui-ci est composé du génome cœur (en anglais : *core genome*), ensemble de gènes partagé par l'ensemble des souches constituant l'espèce, et par le génome dispensable (en anglais : *dispensable genome*), ensemble des gènes présents dans certaines souches mais absents dans certaines (Tettelin et al. 2005 ; Medini et al. 2005), ainsi

que (v) le contenu en G+C (Sueoka 1961 ; Muto et Osawa 1987). Ces techniques seront décrites plus en détail dans la partie A.1.5.2, page 15. Les phylogénies obtenues à partir de ces différentes méthodes n'aboutissent pas au même résultat. C'est donc dans cette incertitude que le quatrième concept d'espèce a vu le jour.

Le concept nominaliste d'espèce Le nominalisme est la « doctrine soutenant que les faits, les lois et les théories scientifiques ne sont que des constructions mentales nécessairement conventionnelles, mais empiriquement fécondes » (Jerphagnon 1973). Le concept nominaliste de l'espèce bactérienne implique que l'espèce bactérienne n'existe pas en tant que tel dans le vivant et que sa description est subjective. Celle-ci est nécessaire et permet de progresser dans la connaissance des bactéries. Erko Stackebrandt, ancien rédacteur du journal *International Journal of Systematic Bacteriology* (IJSB) maintenant connu sous le nom de *International Journal of Systematic and Evolutionary Microbiology* (IJSEM) , déclare dans *The Prokaryotes* : « L'inexistence d'espèces comme catégorie objective et comme produit de la sélection naturelle (...) est reconnue par les microbiologistes depuis plus de 20 ans. Les microbiologistes, en particulier, suivent des principes et des recommandations pour assurer la stabilité, la reproductibilité et la cohérence de la taxonomie — Bien que dans l'analyse finale, la description de l'espèce est toujours subjective » . Le concept de l'espèce nominaliste bactérienne est donc un groupement d'organismes permettant aux microbiologistes de raisonner, d'expérimenter et de communiquer entre eux, mais n'existe pas dans la nature.

Selon Rosselló-Mora et Amann (2001), le « concept d'espèce phylophénétique » qui est une espèce décrite comme « un groupe monophylétique et génomiquement cohérent d'organismes dont de nombreuses caractéristiques indépendantes ont un degré élevé de similarité générale, et qui peuvent être identifiés par un caractère phénotypique discriminant ». Ce concept permet de réunir des organismes dans un taxon qui est réel, mais celui-ci ne peut porter le nom d'espèce au sens strict du terme. Cette définition a été complétée en raison de l'avancée des techniques de séquençage haut-débit : « une catégorie qui circonscrit des populations monophylétiques, génomiques et phénotypiques cohérentes d'individus qui peuvent être clairement distinguées d'autres entités de ce type au moyen de paramètres normalisés » (Rosselló-Móra et Amann 2015). Il est même défendu que chaque espèce bactérienne est unique en son genre et doit être définie par ses propres limites (Palmer et al. 2019).

A.1.3 La définition de l'espèce bactérienne

Bien que le concept d'espèce bactérienne fasse débat, il a été nécessaire d'établir une définition de l'espèce bactérienne, afin de pouvoir classifier les organismes connus et développer une taxonomie bactérienne.

La définition de l'espèce bactérienne selon la commission de Wayne et al. (1987)

En 1987, un comité *ad hoc* de l'International Committee for Systematic of Prokaryotes s'est réuni pour donner une définition phylogénétique de l'espèce bactérienne. Il a été convenu que le premier critère pour définir si différentes souches forment une même espèce est un seuil d'hybridation ADN/ADN d'environ 70% ou plus, et d'une valeur de ΔTm de 5°C maximum. Selon cette définition, à la suite de l'application de ce critère, les caractères phénotypiques de l'espèce doivent correspondre à cette définition et ne doivent outrepasser ce critère que dans de rares cas. Des espèces différentes génétiquement (on peut parler d'espèce génomique, *genospecies* en anglais) et présentant les mêmes caractéristiques phénotypiques ne doivent pas être nommées tant qu'un caractère phénotypique les discriminant n'est pas trouvé. Il est intéressant de constater que ce comité relève l'importance future du potentiel phylogénétique des ARNs ribosomaux, ainsi que la possibilité de séquencer l'intégralité d'un génome bactérien malgré les limites techniques de l'époque.

En raison des avancées techniques, un second comité s'est réuni afin de réévaluer l'espèce bactérienne.

La définition de l'espèce bactérienne selon la commission de Stackebrandt et al. (2002)

Ce comité décide de maintenir le critère d'hybridation ADN/ADN précédemment défini pour définir une espèce, et publie une série de recommandations afin d'intégrer les nouvelles techniques génomiques dans la déclaration d'une nouvelle espèce : (i) il convient d'encourager des techniques génomiques innovantes pouvant compléter ou supplanter l'hybridation ADN/ADN. Le comité cite notamment le séquençage de gènes de ménage ou tout autre gène conservé, l'empreinte génétique ou encore les puces à ADN, (ii) les espèces doivent être identifiables par les techniques les plus simples possibles, (iii) les descriptions des espèces doivent être uniformisées, (iv) la séquence du gène de l'ARNr 16S d'une taille supérieure à 1 300 nucléotides et avec un taux d'incertitude de moins de 0.5% doit être soumise dans la description de l'espèce, (v) l'indication du contenu en G+C de la souche type de l'espèce type d'un nouveau genre doit être incluse dans la description, (vi) il est encouragé de

soumettre plus d'une souche pour toute description d'une nouvelle espèce, et (vii) l'utilisation de la terminologie *Canditatus* pour les espèces non cultivées est encouragée.

Il est donc encouragé de décrire les espèces de la manière la plus précise possible et avec un maximum d'approches possibles : phénotypique, génomique ou phylogénique. Cette définition de l'espèce est appelée approche polyphasique. Ce mot a été introduit pour la première fois par Colwell (1970) dans sa description du genre *Vibrio*. C'est cette définition de l'espèce qui est aujourd'hui appliquée. Elle permet d'apporter une définition de l'espèce la plus précise possible, les organismes composant l'espèce étant monophylétiques, elle présente une cohérence génomique et phénotypique (Rosselló-Móra et Amann 2015).

A.1.4 La taxonomie procaryote

Le terme taxonomie (du grec *taxis* : placement et *nomos* : loi) à été utilisé pour la première fois en 1813 dans *Théorie élémentaire de la botanique* (Candolle 1813). La définition de ce terme varie selon les auteurs (Wilkins 2011) et est sujet à un débat de terminologie entre taxonomie, taxinomie et taxionomie (Tardieu 2011 ; Mayr 1966). Dans notre cas nous utiliserons les définitions données par Simpson (1961) : la taxonomie est considérée comme l'étude théorique de la classification, y compris ses bases, principes, procédures et règles. Les règles pour nommer les taxons (unités taxonomiques de base) tiennent, elles, de la nomenclature. La systématique se définit comme l'étude scientifique des types et de la diversité des organismes et de toutes les relations entre eux (Gillott 1995).

La classification des êtres vivants est présente dans la nature de l'Homme depuis la nuit des temps. Les premières traces écrites d'une classification sont l'oeuvre d'Aristote dans *Histoire des animaux* et *Parties des animaux*, où il décrit les êtres vivants par leurs attributs, tels que le nombre de pattes ou encore la couleur de leur sang. Carl Linnaeus avec son livre *Systema Naturae* (Linné et Salvius 1758) a permis d'uniformiser les noms donnés aux organismes, notamment grâce à l'introduction du nom binomial. Il est le père de la classification classique. Enfin, Charles Darwin avec son livre *On the Origin of Species by Means of Natural Selection, Or, The Preservation of Favoured Races in the Struggle for Life* (Darwin 1859) a fondé la théorie de l'évolution et a apporté les connaissances nécessaires au développement de la classification phylogénétique et à la représentation de la taxonomie en forme d'arbre.

La taxonomie des procaryotes, du fait de leur première observation en 1676 par Antoni van Leeuwenhoek (Leeuwenhoek 1677), n'a pris forme

qu'à partir du XVIII^{ème} siècle. Ceux-ci étaient alors appelés *animalcula* et étaient considérés comme des animaux étant observables seulement au microscope. A la fin du XVIII^{ème} siècle, Otto Friedrich Müller a été le premier à proposer une classification bactérienne sur des critères morphologiques. On parle alors de classification phénotypique, ne se basant que sur une information réduite et avec des critères subjectifs. Il a distingué deux genres de bactéries : *Monas* et *Vibrio*. Le XIX^{ème} et les deux premières décennies du XX^{ème} siècle ont vu l'augmentation du nombre de bactéries décrites ainsi que les premières cultures bactériennes. En 1923, est paru le *Bergey's Manual of Determinative Bacteriology* (Bergey 1923) qui est la première clé d'identification des bactéries. Il n'y avait pas à l'époque de concertation dans la communauté scientifique sur la classification des procaryotes et les erreurs de nomenclature étaient fréquentes. Le *Bergey's Manual of Determinative Bacteriology* devient la référence en matière de classification des procaryotes. A la fin des années 1950 est apparue la taxonomie numérique (Sneath 2015), qui via l'utilisation d'algorithmes numériques (Rogers et Tanimoto 1960), permet d'agréger l'ensemble des états de caractères phénotypiques et la formation de groupes de bactéries similaires via le calcul d'un indice numérique (l'indice et le coefficient de Jaccard). Celui-ci évalue les similitudes phénétiques des souches d'intérêts et calcul une distance taxonomique entre chacune d'entre-elles. Cette dernière permet d'obtenir des groupes homogènes d'organismes que l'on peut appeler espèce, et d'organiser ces espèces en genre et en rang de plus haut niveau. Grâce au développement des techniques de biologie moléculaire, cette période a aussi permis l'émergence de la chimiotaxonomie qui a pour objectif d'identifier et classer les organismes par les différences et similitudes de leur composition biochimique. Les éléments chimiques d'intérêt sont entre autres les protéines, les sucres ou encore les lipides constitutifs de la membrane (Vandamme et al. 1996).

Mais le début des années 60 marque aussi l'approfondissement des connaissances sur l'ADN et fait apparaître des méthodes tirant profit de ces avancées. Des approches génétiques sont utilisées afin d'identifier et de classer les souches procaryotes. L'hybridation ADN/ADN est notamment reconnue par la commission de Wayne et al. (1987) comme principal critère pour définir une espèce procaryote. Le contenu en G+C a également permis de comparer différentes souches et leur proximité. Dans les années 70, le séquençage de l'ADN a permis le début des approches phylogénétiques, notamment la reconstruction d'arbre phylogénétique basée sur le gène de l'ARNr 16S. La baisse des coûts et les progrès en matière de séquençage ont permis l'apparition d'autres méthodes phylogéniques, telles que l'identification génétique (génotypage), le typage par séquençage multilocus ou le séquençage entier de génomes.

Ces différentes approches phénotypiques, génétiques et phylogéniques combinées ont permis l'émergence de la taxonomie polyphasique (aussi appelée taxonomie mixte et consensuelle, Vandamme et al. 1996 ; Gevers et al. 2006). Cette approche de la taxonomie permet d'agréger des données provenant de différentes sources, afin d'obtenir une classification consensuelle avec le moins de contradictions possibles. Elle reprend les principes décrits dans la définition de l'espèce polyphasique : la convergence de toutes les informations génotypiques, phénotypiques et phylogénétiques afin de déterminer les taxons bactériens (Zakhia et Lajudie 2006).

A.1.5 Méthodes utilisées pour la classification bactérienne

A.1.5.1 Les méthodes phénotypiques

Les caractères phénotypiques sont les caractéristiques observables qui résultent de l'expression des gènes d'un organisme (Moore et al. 2010). Les premiers caractères phénotypiques utilisés étaient purement morphologiques et caractérisaient l'organisme : taille, forme, présence d'un flagelle, endospore, corps d'inclusion, ou la colonie : couleur, forme, dimension. Ces critères sont dépendants des conditions de culture ou de l'environnement. La composition de la paroi cellulaire est également un caractère phénotypique permettant de discriminer différentes bactéries via notamment la coloration de Gram qui réagit en fonction de la composition de la paroi cellulaire (Schleifer et Kandler 1972). Le sérotypage des cellules via les anticorps (Henriksen 1978), l'analyse des isoprénoïde quinone présentes dans la membrane cellulaire (Collins et Jones 1981) et la comparaison des profils protéiques des individus (Kersters et al. 1994) permettent également de classifier les bactéries. La liste des techniques présentée ici n'est pas exhaustive et il existe des approches propres à certains taxons. Toutefois, l'ensemble de ces méthodes ne permet d'avoir accès à l'expression d'une partie réduite du génome.

A.1.5.2 Les méthodes génotypiques

Les caractéristiques physico-chimiques de l'ADN ont servi de base au développement des premières méthodes permettant de délimiter l'espèce bactérienne.

L'hybridation ADN/ADN repose sur les propriétés de dénaturation et de renaturation de l'ADN. Le chauffage progressif d'une molécule d'ADN entraîne la rupture des liaisons hydrogènes reliant les paires de bases entre les

deux brins. Les brins deviennent indépendants l'un de l'autre et seul un refroidissement lent permet de rétablir ces liaisons hydrogènes. La température de dénaturation est dépendante du contenu en G+C du génome (la température augmente si un génome présente un fort contenu en G+C, Marmur et Doty 1962). L'écart entre la température de renaturation de l'ADN et celle de dénaturation est comprise entre 22 et 26°C.

Afin de déterminer si deux organismes appartiennent à la même espèce, leur ADN est mélangé dans une même solution. Après dénaturation, la température de la solution est abaissée afin que les brins d'ADN se réassocient. La température de renaturation de l'ADN est en moyenne de 24°C inférieure à sa température de dénaturation. Lors de la renaturation on obtient deux types de duplex : (i) des homoduplex d'ADN formés de deux brins d'ADN provenant du même organisme et (ii) des hétéroduplex d'ADN formés de deux brins provenant chacun d'un organisme différent (hybrides). C'est le ratio d'hétéroduplex qui est pris en compte dans le taux d'hybridation ADN/ADN. Plus les brins d'ADN des deux espèces sont complémentaires, plus cela favorisera l'apparition d'hétéroduplex. S'il y a plus de 70% d'hétéroduplex, alors les deux bactéries peuvent être considérées comme étant de la même espèce (Wayne et al. 1987). Il existe plusieurs techniques afin de mesurer le ratio d'hétéroduplex, notamment (i) la méthode à l'hydroxyapatite (Brenner et al. 1969), (ii) la méthode à l'endonucléase S1 (Crosa, Brenner et Falkow 1973) et (iii) la méthode de renaturation optique (Ley, Cattoir et Reynaerts 1970). Les deux premières méthodes permettent la caractérisation de la stabilité thermique de ces hétéroduplex vis-à-vis des homoduplex par le calcul du ΔT_m, c'est-à-dire la différence de température entre la température de dénaturation des homoduplex et celle des hétéroduplex. Si cette différence de température est faible, alors les hybrides formés possèdent peu de bases non appariées et donc en conséquence une séquence d'autant plus proche. Une différence de température de dénaturation entre les homoduplex et les hétéroduplex entre 1 et 2.2% correspond à 1% de bases non appariées dans l'hétéroduplex. Ainsi, en plus de 70% d'hybridation ADN/ADN, deux organismes appartiennent à la même espèce si le ΔT_m n'est pas supérieur à 5°C (Wayne et al. 1987).

Le contenu en G+C d'un génome est également utilisé pour comparer les génomes des procaryotes sur la base des paires d'acides nucléiques G-C est un prérequis pour déterminer les conditions de températures devant être utilisées lors de l'hybridation ADN/ADN. Ce pourcentage était dans un premier temps estimé à partir de méthodes indirectes (c.-à-d. qu'elles ne font qu'estimer le contenu, elles ne permettent pas d'obtenir une valeur absolue)

basées sur la température de dénaturation de l'ADN (Marmur et Doty 1962), la centrifugation à l'équilibre dans une solution de chlorure de césium (Schildkraut, Marmur et Doty 1962), des températures de fusion de l'ADN (Owen, Hill et Lapage 1969) ou encore par des méthodes plus récentes comme la PCR quantitative (Moreira, Pereira et Thompson 2011) ou la chromatographie (Ko et al. 1977). Les avancées en matière de séquençage de génome complet permettent la détermination précise du %G+C grâce à l'accès à la séquence des acides nucléiques et via la formule suivante (Rosselló-Mora et Amann 2001) :

$$\%G + C = [G + C]/[A + T + C + G] \times 100$$

Le %G+C varie entre 13 et 76% chez les procaryotes (Vandamme et al. 1996 ; McCutcheon et Moran 2010) et le %G+C des organismes d'une même espèce bactérienne ne doivent pas différer de plus de 3% ou 5% selon les définitions de l'espèce (Mesbah, Whitman et Mesbah 2011 ; Rosselló-Mora et Amann 2001) et de plus de 10% dans le cas d'un genre.

Ces techniques sont toujours utilisées de nos jours (Meier-Kolthoff, Klenk et Göker 2014) le %G+C devant être soumis dans la description d'une nouvelle espèce et l'hybridation ADN/ADN étant encore la méthode de référence pour définir une espèce bactérienne (Tindall et al. 2010). Mais cette dernière est très lourde à mettre en œuvre, notamment pour les organismes ne pouvant être cultivés. Le deuxième comité de définition de l'espèce bactérienne (Stackebrandt et al. 2002) encourage à compléter l'hybridation ADN/ADN par de nouvelles méthodes quantitatives pouvant être validées sur de nombreuses collections de souches. Si une méthode suivant ces principes se dégage, alors elle pourrait remplacer l'hybridation ADN/ADN comme technique de référence.

Le gène de l'ARNr 16S est le gène qui code l'ARN ribosomal 16S. L'ARNr 16S forme, avec un ensemble de protéines, la petite sous-unité 30S du ribosome bactérien. Il a été l'un des premiers gènes marqueurs utilisés dans la classification des bactéries (Woese 1987), car celui-ci est présent chez les bactéries et les archées. Ce gène présente une taille relativement constante (1 378 nucléotides valeur médiane avec un écart-type de 57 nucléotides avec des valeurs minimales et maximales de 1 112 et 2 369 nucléotides respectivement. Cette valeur est calculée sur l'intégralité des séquences complètes des gènes de l'ARNr 16S présente dans GreenGenes 13.5, McDonald et al. 2012). Il est constitué de neuf régions variables entourées de régions conservées (Figure A.III).

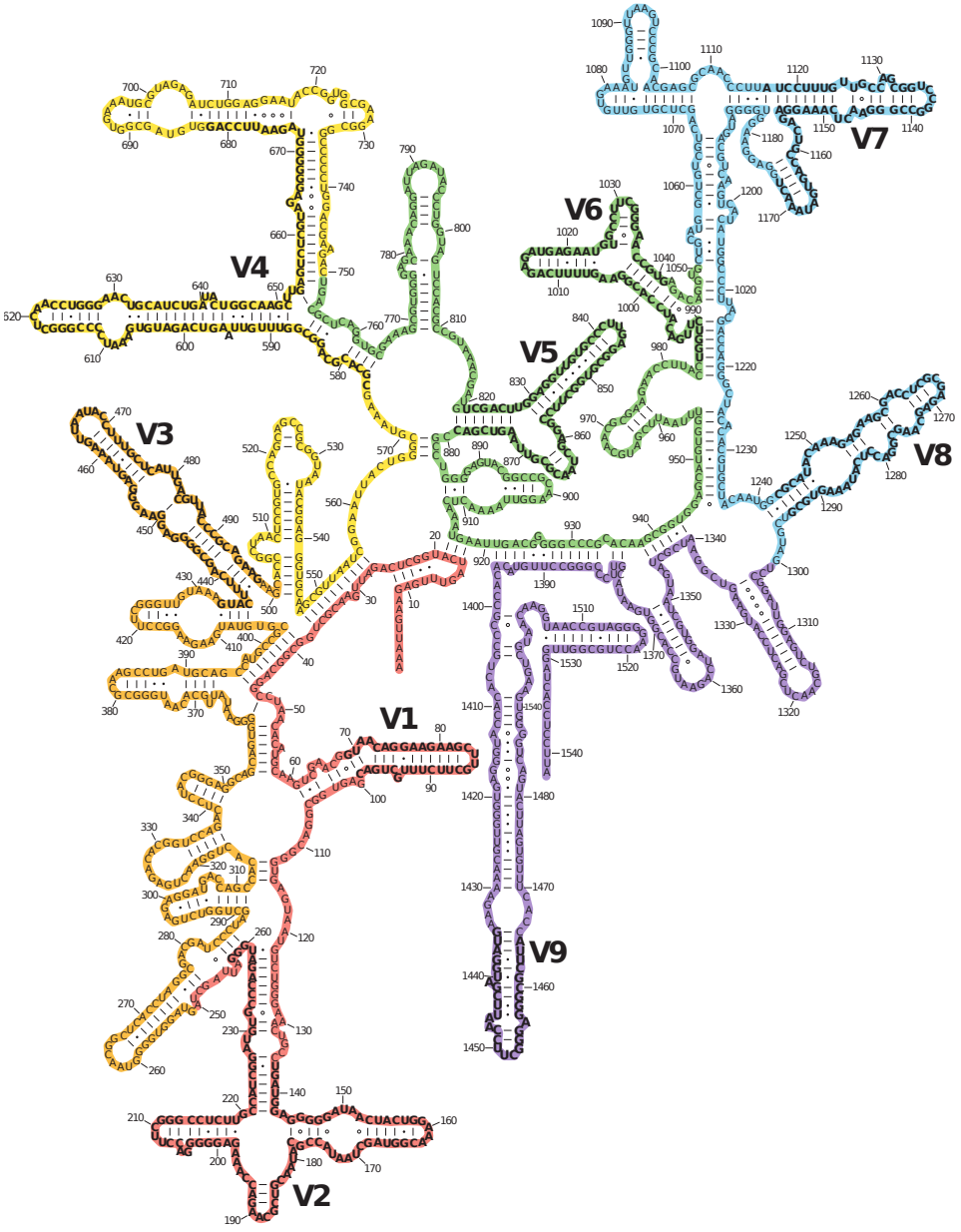


FIGURE A.III – Structure secondaire de l'ARNr 16S chez *Escherichia coli* composée de 1541 nucléotides (BROSIUS et al. 1978). Les régions variables sont indiquées par leur nom (**V1** par exemple), et les nucléotides les composant sont indiqués en gras.

Source: YARZA et al. (2014)

Ces régions variables ont des tailles comprises entre une quarantaine et une centaine de paires de bases et sont nommées de V1 à V9 en fonction de leur position (Chakravorty et al. 2007). Ces régions variables sont par définition sujet à des substitutions, des insertions ou à des délétions de nucléotides tout en conservant la structure secondaire de l'ARNr. Les différentes régions conservées sont également très facilement amplifiables par PCR grâce à des amorces universelles (Weisburg et al. 1991). La principale limite du gène de l'ARNr 16S est qu'il est en multiples copies dans la majorité des génomes. En effet, *Photobacterium damselae* possède 21 copies (Stoddard et al. 2015) de ce gène. Le nombre moyen de copies par organisme chez les bactéries oscille entre 4,2 copies (Větrovský et Baldrian 2013) et 4,8 copies (selon la base de données rrnDB (Ribosomal RNA Database), Stoddard et al. 2015). Seules environ 15% des bactéries ne contiendraient qu'une seule copie. Pour les archées, le nombre maximal de copies est de 4 pour une moyenne de 1,7 copies. Au sein d'un organisme les copies du gène de l'ARNr 16S n'ont pas forcément la même séquence. Des souches différentes présentant des séquences du gène de l'ARN 16S différentes sont qualifiées de ribotype (Case et al. 2007), chacun de ces ribotypes pouvant présenter un génotype différent (Thompson et al. 2005). Ces différentes copies pour un même organisme peuvent mener à des problèmes de phylogénie ou de taxonomie (Marchandin et al. 2003). De plus, Kitahara et Miyazaki (2013) ont montré que le gène de l'ARNr 16S pouvait être occasionnellement sujet à un transfert horizontal entre organismes procaryotes ce qui rend, par conséquent, la classification des organismes plus difficile.

La commission de Stackebrandt et al. (2002) reconnaît l'utilité du gène de l'ARNr 16S dans le cadre d'une taxonomie bactérienne et impose le dépôt de la séquence du gène lors de chaque description d'une nouvelle espèce bactérienne. Le seuil de 97% de similarité entre deux séquences du gène de l'ARNr 16S avait été alors retenu (Stackebrandt et Goebel 1994) ce qui correspond à une différence maximum de 45 nucléotides entre les gènes. Ce seuil a été choisi suite au constat que deux génomes bactériens ne peuvent avoir plus de 70% d'hybridation ADN/ADN que si, et seulement si, la similarité de leur séquence du gène de l'ARNr 16S est supérieure à 97% (Figure A.IV). Ce seuil permet d'obtenir une première indication sur la possible appartenance ou non de deux organismes à une même espèce sans faire appel à l'hybridation ADN/ADN. Si deux bactéries présentent des séquences du gène de l'ARNr 16S ayant une similarité inférieure à 97% alors elles ne peuvent être dans la même espèce et il n'est pas nécessaire de réaliser une hybridation ADN/ADN des génomes. Ce n'est que si les gènes présentent plus de 97% de similarité de séquence qu'il est alors nécessaire de faire appel à l'hybridation ADN/ADN afin de confirmer un taux d'hybridation supérieur ou égal à

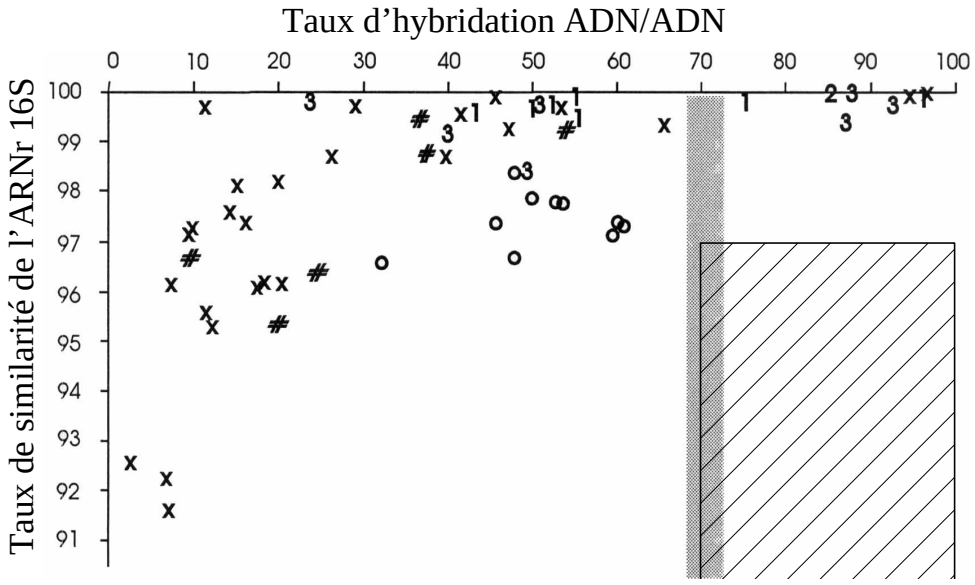


FIGURE A.IV – Comparaison du taux de similarité de l'ARNr 16S et du taux d'hybridation ADN/ADN dans une sélection d'organismes provenant de différents jeux de données. Chaque symbole indique un jeu de données différent où l'hybridation ADN/ADN à été effectué via différentes méthodes : **x**, méthode par filtration membranaire ; #, 1, 2, 3 et 4, méthode par taux de renaturation ; ∘ , méthode de la nucléase S1. La zone hachurée (moins de 97% de similarité entre les ARNr 16S et plus de 70% d'hybridation ADN/ADN) ne contient aucun couple d'organismes.

D'après STACKEBRANDT et GOEBEL (1994)

70% entre les deux organismes et donc de valider que ceux-ci appartiennent à la même espèce. Cette valeur de similarité permet d'éviter l'hybridation ADN/ADN, technique lourde et peu reproductible mais nécessaire à la définition de l'espèce bactérienne. Ce seuil de 97% est appelé l'étalon-or (de l'anglais : *gold standard*) de l'identification bactérienne.

Il a été proposé en 2006 de modifier ce seuil pour des raisons de reproductibilité des résultats et de praticité (les taxonomistes n'étant pas forcément enclins à réaliser des hybridations ADN/ADN). En effet, l'hybridation ADN/ADN est difficilement reproductible par un pair car dépendant de nombreux facteurs physico-chimiques alors que le séquençage du génome l'est davantage. La proposition est de réévaluer ce seuil entre 98,7-99% (soit entre 15 et 20 nucléotides de différence entre gènes de l'ARNr 16S) afin de rendre plus discriminante cette technique (Figure A.V) et ainsi avoir moins recours à l'hybridation ADN/ADN (STACKEBRANDT et EBERS 2006). En dessous

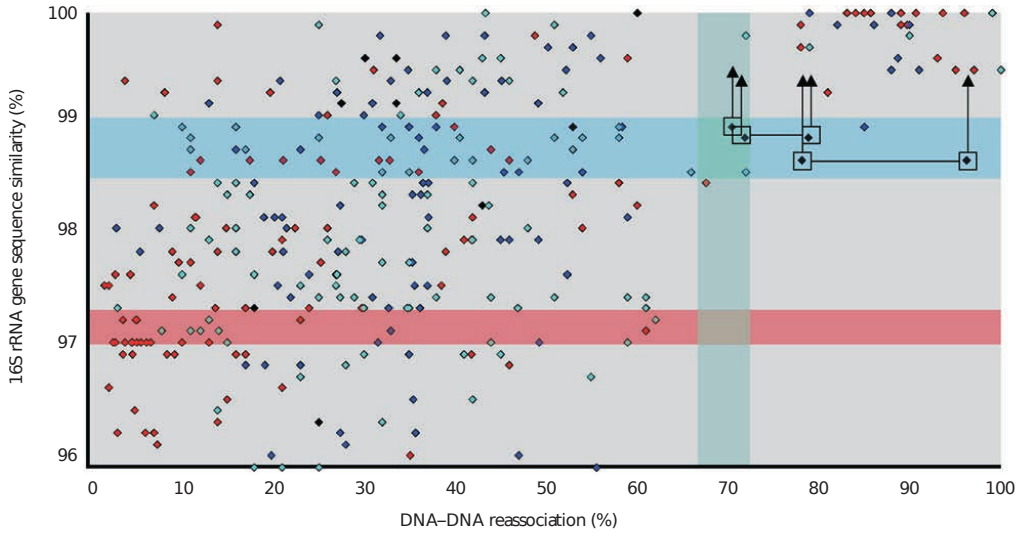


FIGURE A.V – Comparaison des taux de similarité de l'ARNr 16S et du taux d'hybridation ADN/ADN. Chaque symbole indique un jeu de données différents provenant d'IJSME Volume 55 (2005). La forme du symbole indique la technique de renaturation. La barre verticale rouge indique l'ancien seuil défini dans STACKEBRANDT et GOEBEL (1994) tandis que la barre bleue indique le seuil de 98,7-99% proposé. Les barres horizontales entre les carrés indiquent les résultats obtenus entre différentes méthodes d'hybridation ADN/ADN tandis que les flèches horizontales indiquent le recalcul *in silico* de la similarité de l'ARNr 16S pour les organismes concernés.

Source: STACKEBRANDT et EBERS (2006)

de ce seuil, l'appartenance à une même espèce est à démontrer par d'autres techniques.

Le gène de l'ARNr 16S s'imposant comme l'étalon-or de l'identification et la classification d'espèce bactérienne, il a été nécessaire de mettre en place des banques de données de séquences. C'est dans ce cadre que naissent de nombreuses bases de données dont *The Ribosomal Database Project* (RDP), OLSEN et al. 1992 ; COLE et al. 2014). Cette base de données a été construite en réunissant l'intégralité des séquences de précédentes collections publiques ou privées, de GenBank, d'EMBL. Ces séquences ont ensuite été alignées et proposées à la libre consultation, associées avec des outils permettant des analyses et des comparaisons entre séquences. Aujourd'hui, la plupart des séquences proviennent de l'*International Nucleotide Sequence Database Collaboration* (INSDC) qui est le fruit de la collaboration du DDBJ, EMBL-

Rang	Seuil	Minimum (%)	Mediane (%)
Espèces	98,7	98,7	NA
Genre	94,5	94,8 (94,5, 95,1)	96,4 (96,2, 96,6)
Famille	86,5	87,7 (86,8, 88,4)	92,3 (91,7, 92,9)
Ordre	82	83,6 (82,3, 84,8)	89,2 (88,3, 90,1)
Classe	78,5	80,4 (78,6, 82,5)	86,4 (84,7, 88,0)
Phylum	75	77,5 (75,0, 79,9)	83,7 (81,6, 86,0)

Tableau A.2 – Seuils de similarité de séquence du gène de l'ARNr 16S pour les rangs taxonomiques majeurs. L'intervalle de confiance est indiqué entre parenthèses. Le seuil de qualification de l'espèce a été déterminé par Stackebrandt et Ebers (2006).

Source: Rosselló-Móra et Amann (2015) et Yarza et al. (2014)

EBI et NCBI. Toutes les séquences sont affiliées à une taxonomie respectant le *Bergey's Manual Trust* (`https://www.bergeys.org/`) ainsi que *The List of Prokaryotic Names with Standing in Nomenclature* (LPSN, `http://www.bacterio.net/`, Parte 2018). La base de données Greengenes (DeSantis et al. 2006 ; McDonald et al. 2012, voir partie A.1.7.5, page 36) contient également ce type de séquence, mais celle-ci ne semble plus mise à jour depuis 2013. Le projet LTP (Living Tree Project, Partie A.1.7.3, page 34) réunissant plusieurs partenaires (ARB, SILVA, LPSN et *the journal Systematic and Applied Microbiology*) a pour but de reconstruire un arbre phylogénétique basé sur l'ARNr 16s de toutes les souches type de procaryotes (Yarza et al. 2008). Les auteurs de la base de données SILVA (Quast et al. 2013) tirent parti de cet effort de curation afin de proposer des données de grande qualité et mettent à disposition une taxonomie qui leur est propre (Yilmaz et al. 2014). Cette taxonomie a été dernièrement réorganisée en suivant le projet « Genome Taxonomy Database » (GTDB, Partie A.1.7.8, page 37) qui a permis une révision de l'arbre du vivant (Parks et al. 2018 ; Waite et al. 2017). Le projet EzTaxon (Chun et al. 2007) suivi du projet EzBioCloud (Yoon et al. 2017a) propose une approche similaire permettant l'identification d'organismes par comparaison avec une base de données d'ARNr 16S.

Grâce à ces données, il a été possible de proposer de nouvelles règles afin de limiter les rangs supérieurs à celui de l'espèce (Tableau A.2). Ces rangs, bien qu'abstraits et que leur cohérence soit inversement corrélée à leur rang, présentent des seuils de similarité de la séquence du gène de l'ARNr 16S (Yarza et al. 2014).

Mais l'identification via le gène de l'ARNr 16S peut ne pas être assez discriminante (Fox, Wisotzkey et Jurtshuk 1992 ; Mignard et Flandrois

2006). De plus, de nombreux biais entrent en compte (différentes copies, possible transfert horizontal d'une partie du gène, biais technique lors du séquençage). L'utilisation d'autres gènes marqueurs de diversité pourrait constituer une solution additionnelle à ces limites.

Les autres gènes marqueurs de la diversité. Les avancées en matière de séquençage ont permis de séquencer des génomes procaryotes entiers. La comparaison de ces génomes a mis en évidence l'existence d'un génome cœur défini comme l'ensemble des gènes présents dans l'intégralité des organismes d'un clade d'intérêt. Cette notion peut s'appliquer à une espèce bactérienne ou à l'intégralité des procaryotes (le gène de l'ARN 16S fait partie du génome cœur des procaryotes). Une partie des gènes composant le génome cœur sont des gènes dits de ménage (en anglais : *housekeeping genes*) assurant les fonctions de base de la cellule et codant pour des protéines ubiquitaires aux procaryotes (Sarkar et Guttman 2004). Les gènes marqueurs généralement choisis sont : (i) les gènes codant sur les sous-unités α et β de l'ADN polymérase (*rpoA* et *rpoB*, ce dernier étant présent en une seule copie dans chaque organisme, Case et al. 2007), (ii) le gène codant pour la sous-unité β de l'ADN gyrase (*gyrB*) ou encore (iii) *recA*, le gène de la recombinase A.

Aucun de ces gènes n'a encore démontré son pouvoir discriminant à travers les différents phyla bactériens et, bien que les premiers résultats soient encourageants, il est encore nécessaire de les tester dans l'intégralité des taxons bactériens (Poirier et al. 2018).

Bien que le comité de Stackebrandt et al. (2002) pensait que le séquençage d'autre gènes marqueurs puisse remplacer l'hybridation ADN/ADN lors de la délimitation d'une espèce, le choix d'un gène marqueur additionnel au gène de l'ARNr 16S ainsi que le seuil de délimitation d'une espèce n'ont jamais pu être définis (Glaeser et Kämpfer 2015).

L'analyse de séquences multilocus (MLSA de l'anglais : *Multi-Locus Sequence Analysis*) est une technique reposant sur ces gènes marqueurs permettant la reconstruction de l'histoire évolutive des organismes, afin d'en déduire les relations de parenté et, *in fine*, une classification bactérienne. Celle-ci est dérivée de la technique du typage par analyse de séquence (MLST de l'anglais : *Multi-Locus Sequence Typing*) permettant d'identifier des souches bactériennes en épidémiologie en se reposant sur les différences alléliques de différents gènes entre différentes souches (Maiden et al. 1998). La MLSA reprend ce principe mais avec des gènes de ménage (Gevers et al. 2005). Cette technique nécessite de sélectionner le plus de gènes possibles et au minimum quatre ou cinq (Glaeser et Kämpfer 2015) pour des organismes

proches. Si les organismes sont relativement distants, alors il est nécessaire de faire appel à un plus grand nombre de gènes ce qui permet de surmonter les problèmes de bruit phylogénétique (Rosselló-Móra et Amann 2015). Les critères de choix des gènes sont les suivants (De Vos 2011) : (i) ils doivent tous être présents dans tous les organismes à discriminer, (ii) le gène doit être présent en une seule copie dans chacun des génomes étudiés, (iii) les différents gènes ne doivent pas avoir subi de transferts horizontaux, (iv) la variabilité de la séquence des gènes doit coïncider avec la variabilité des génomes entiers, et (v) des amorces [2] doivent être disponibles ou peuvent être synthétisées afin de pouvoir amplifier les gènes d'intérêt en vue de leur séquençage. Une fois les gènes choisis, ils sont amplifiés par PCR et séquencés. Les lectures obtenues d'un gène d'un organisme sont ensuite alignées contre les lectures du même gène chez les autres organismes à discriminer. On parle d'alignement multiple de séquences (en anglais : MSA : *Multiple Sequence Alignment*). On obtient donc autant d'alignements multiples de séquences qu'il y a de gènes d'intérêt. Ces alignements multiples de gènes sont ensuite concaténés afin de reconstruire un arbre phylogénétique consensus pour tous les gènes. Les techniques d'inférence phylogénétique et de reconstruction d'arbres phylogénétiques et d'arbres phylogénétiques consensus seront abordées dans la partie A.1.5.3, page 26. Cet arbre consensus permet d'identifier les souches appartenant à un même clade et donc à la même espèce ou genre. Mais comme pour les gènes marqueurs uniques, il n'existe pas de consensus sur le choix des gènes, ainsi que sur une valeur seuil pour la définition de l'espèce bactérienne.

Le séquençage génome complet et les indices globaux de parenté génomique. Le séquençage de génome complet étant facilité par les avancées techniques, il est maintenant possible de réaliser des analyses sur la séquence complète du génome. Ces analyses rendent possible le calcul d'indices globaux de parenté génomique (en anglais : OGRI (*Overall Genome Relatedness Indices*, Chun et Rainey 2014). Ces indices ont pour but de supplanter l'hybridation ADN/ADN dans le processus de délimitation d'espèce bactérienne car l'hybridation ADN/ADN est difficilement répétable par les paires alors qu'une fois le génome séquencé et sa séquence déposée dans une banque de données publique telle que Sequence Read Archive (SRA), il est alors possible de répéter le calcul des OGRI sur la séquence facilement. Il existe plusieurs OGRI : (i) l'ANI (identité moyenne de la séquence nucléotidique, de l'anglais : *Average Nucleotide sequence Identity*, Konstantinidis

2. Courtes séquences se plaçant par hybridation au début et à la fin du gène d'intérêt, elles permettent la délimitation du gène et servent de point de départ à l'ADN polymérase lors de la PCR.

et Tiedje 2005a ; Goris et al. 2007 ; Arahal 2014, (ii) la dDDH (hybridation ADN/ADN numérique, de l'anglais : *digital DNA/DNA hybridation*, (Auch et al. 2010) et (iii) le MUM (Nombre maximal de correspondances uniques, de l'anglais : *Maximal Unique Matches*, Deloger, El Karoui et Petit 2009).

Parmi ces trois indices, l'ANI est le plus utilisé en substitut *in silico* de l'hybridation ADN/ADN pour l'appartenance ou non à une même espèce bactérienne. La détermination de l'ANI est réalisée en deux étapes. Dans un premier temps, l'intégralité des génomes des organismes comparés sont alignés entre eux et les régions correspondantes sont identifiées. Dans un deuxième temps, le calcul de l'identité nucléotidique est effectué à partir de ces régions correspondantes. L'ANI correspond à la moyenne des valeurs d'identité nucléotidique. Différentes méthodes d'alignement et différentes données d'entrée peuvent être utilisées pour calculer l'ANI. Les principales méthodes sont : (i) l'ANIb, méthode où un des génomes est coupé en fragments de 1 020 nucléotides et chacun de ces fragments est recherché dans le génome complet de l'autre organisme (Goris et al. 2007), (ii) l'ANIm, qui repose sur le logiciel MUMmer (Kurtz et al. 2004) où les deux génomes comparés ne sont pas fragmentés (Richter et Rosselló-Móra 2009), (iii) OrthoANI, où les deux génomes sont fragmentés en séquences de 1 020 nucléotides et seul les fragments présentant une orthologie sont utilisés dans le calcul (Lee et al. 2016), et (iv) FastANI, ne reposant pas sur des alignements de séquences mais sur une estimation de l'identité des k-mers (Jain et al. 2018b ; Jain et al. 2018a).

L'indice ANI est exprimé en pourcentage d'identité. Une valeur d'hybridation ADN/ADN supérieure à 70% correspond à un ANI supérieur ou égal à 94%. Rosselló-Móra et Amann (2015) définit ce seuil à 96% car entre 93% et 96% il existe une incertitude. L'ANI est fondé sur la séquence du génome complet. Il est donc nécessaire, pour calculer l'ANI dans le cadre d'étude métagénomique d'environnement complexe, de séquencer avec la plus grande qualité et profondeur possible. Il est recommandé d'effectuer un assemblage présentant des séquences de grandes tailles pour calculer l'ANI (Jain et al. 2018b ; Chun et al. 2018). On peut notamment se baser sur l'indicateur N50 qui est la taille du scaffold [3] tel que 50% de la longueur cumulée de l'ensemble des scaffolds (ou contigs) obtenus après assemblage soit contenue dans des scaffolds (ou contigs) de taille égale ou supérieure (Bruto 2010). A noter que l'ANI n'est pas un bon indicateur pour les organismes très distants

3. Ensemble de contigs ordonnés et orientés. Un scaffold contient des parties de séquences où les nucléotides ne sont pas connus, mais malgré ces interstices il existe des preuves supportant l'ordre et l'orientation des contigs composant le scaffold.

(présentant un ANI inférieur à 80%). Il est recommandé dans ce cas de se baser sur l'AAI (Moyenne d'identité des acides aminés, en anglais : *Average Amino acid Identity*, Konstantinidis et Tiedje 2005b ; Rodriguez-R et Konstantinidis 2014).

Les indices OGRI pourraient remplacer l'hybridation ADN/ADN pour déterminer si une souche séquencée appartient à une espèce déjà connue. Elle permet aussi de délimiter de nouvelles espèces bactériennes. Il est actuellement proposé que ces indices, accompagnés d'un maximum d'information sur le séquençage et de statistiques sur l'assemblage, soient exigés pour la déclaration de nouvelles espèces (Chun et al. 2018). Cela a pour but de faciliter la reproductibilité des analyses taxonomiques et de faciliter les approches taxonomiques futures.

Les seuils des différentes méthodes génotypiques évoquées dans cette partie sont résumés dans le tableau A.3.

Méthodes	**Seuil**	**Inconvénients**
Hybridation ADN/ADN	70%	Peu reproductible, lourd techniquement
ARNr 16S	98,7%	Définition limitée au genre, ribotype et possibilité de transfert horizontal
Autre gènes marqueurs (*gyrB*, *rpoA*, *rpoB*, *recA*), multilocus sequence analysis	Non consensuel	Choix du/des gène(s) et du seuil intrinsèque au clade d'intérêt
ANI	93-96%	Dépendant de la qualité de séquençage et d'assemblage

Tableau A.3 – Seuil de définition de l'espèce associé aux différentes méthodologies.

A.1.5.3 L'apport de la phylogénie

Dans les différentes techniques vues précédemment (ARNr 16S, MLSA), ainsi que dans la définition de l'espèce phylogénétique (Partie A.1.2, page 10), il est intéressant de considérer plusieurs organismes en même temps, et de ne pas se limiter à une comparaison organisme à organisme. Pour cela, il est possible de générer des arbres phylogénétiques qui représentent les relations évolutives entre différents organismes en se basant sur leurs similarités ou leurs différences. Chaque branche représente la persistance d'une lignée génétique

à travers le temps et chaque noeud représente la naissance d'une nouvelle lignée (Yang et Rannala 2012). Un clade est défini comme l'intégralité des organismes partageant un même ancêtre commun. On peut également parler de groupe monophylétique. Les arbres phylogénétiques ne peuvent être directement observés et doivent être reconstruits à partir de séquences nucléotidiques telles que le gène de l'ARNr 16S ou de séquences protéiques de gènes marqueurs. L'inférence de l'arbre phylogénétique fait donc appel à des méthodes de reconstruction d'arbres phylogénétiques. Il existe trois méthodes principales, elles-mêmes pouvant inclure des sous-méthodes :

Méthode basée sur une matrice de distance Dans un premier temps, une distance est calculée entre paires de séquences grâce à leur alignement. Chacune de ces distances est calculée grâce à un modèle d'évolutions des séquences d'ADN, des codons, ou des acides aminés (Arenas 2015). Une matrice de distance est obtenue. Un arbre est ensuite généré grâce à plusieurs méthodes, telle que la méthode des moindres carrés, de l'évolution minimum, ou UPGMA (Unweighted pair group method with arithmetic mean). La plus utilisée est toutefois celle du « Neighbor Joining » (Saitou et Nei 1987).

Méthode basée sur les caractères Il existe deux principales méthodes basées sur les caractères : (i) le maximum de parcimonie, où le nombre de changements d'état de caractères (autrement dit, de changement de nucléotide pour un site donné dans la séquence) doit être réduit au maximum. Le meilleur arbre est celui qui implique le moins de changements d'état de caractère. (ii) Le maximum de vraisemblance, qui repose également sur un modèle d'évolution tel que décrit précédemment, ainsi que sur la loi des probabilités conditionnelles. Il permet d'obtenir l'arbre ayant la plus forte probabilité de représenter l'évolution des séquences conduisant aux séquences sources.

Méthode bayésienne La méthode bayésienne est une méthode statistique permettant de calculer *a posteriori* la probabilité qu'un arbre soit vrai, sachant les données que l'on possède. Cette méthode demandant une puissance de calcul très importante, elle est combinée avec l'utilisation de chaînes de Markov avec technique de Monte Carlo, qui permet de guider la méthode à travers l'ensemble des arbres possibles.

L'analyse phylogénétique de larges jeux de données Avec les avancées en matières de séquençage, il est désormais nécessaire que les méthodes d'inférence phylogénétique puissent gérer des données très importantes en terme de taille. C'est le cas, par exemple, pour la MLSA (Partie A.1.5.2,

page 23), où les relations phylogénétiques de plusieurs gènes marqueurs sont considérées. Il est donc question d'assembler ces relations phylogénétiques, afin de composer un arbre final des relations phylogénétiques. Pour cela il existe deux méthodes (Sanderson, Purvis et Henze 1998) : (i) l'approche du « super arbre », qui assemble chacun des sous arbres définis pour chaque gène en un seul arbre et (ii) l'approche « super matrice », qui concatène l'ensemble des alignements de séquences et ensuite trace un arbre phylogénétique.

La phylogénie joue un rôle prépondérant dans la taxonomie polyphasique (Partie A.1.4, page 14), mais celle-ci doit être complétée par d'autres méthodes lors de la description d'un nouveau taxon.

A.1.6 Nomenclature bactérienne

Dans les précédentes parties nous avons décrit les critères permettant d'affilier une souche à l'espèce ou à un autre rang taxonomique. Néanmoins, afin de faciliter la communication entre chercheurs, il est nécessaire d'établir des règles concernant les noms des procaryotes. Les procaryotes possèdent leur propre code intitulé *International Code of Nomenclature of Prokaryotes* ou ICNP (Parker, Tindall et Garrity 2019). La première pré-version a été publiée en 1947 (Buchanan, St. John-Brooks et International Congress for Microbiology 1947), mais la première publication n'a lieu que 28 ans plus tard en 1975. Depuis, des révisions ont été validées en 1990 et publiées en 1992 (Lapage et al. 1992) ainsi que proposées en 2008, présentées en 2014 et publiées en 2019 (Parker, Tindall et Garrity 2019). Avant la publication du premier code de nomenclature des procaryotes, la nomenclature de ceux-ci était soumise au code *International Code of Nomenclature for algae, fungi, and plants* (ICN, anciennement *International Code of Botanical Nomenclature*, Turland et al. 2018) qui gère la nomenclature des plantes, algues et champignons. Les cyanobactéries, unique phylum contenant des bactéries étant capables de réaliser la photosynthèse afin d'obtenir de l'énergie et longtemps considérées comme des plantes unicellulaires sont encore soumises aux règles de nomenclature l'ICN en parallèle de celle de l'ICNP (Oren 2011 ; Oren et Ventura 2017), bien que des efforts soient faits pour trouver un consensus (Palinska et Surosz 2014 ; Pinevich 2015).

L'ICNP définit l'ensemble des principes et des règles qui régulent les noms des taxons. Les principes de bases sont les suivants : (i) le principal objectif est la stabilité des noms dans le temps. Les noms pouvant causer des erreurs ou des confusions seront rejetés et la création de noms inutiles doit être proscrite. (ii) Les noms des rangs supérieurs ou égaux au genre (tels que par exemple le nom des ordres et des familles) ne doivent pas être présents

dans l'*International Code of Nomenclature for algae, fungi, and plants* et dans l'*International Commission on Zoological Nomenclature* (code régulant la nomenclature du domaine des animaux). Cette règle ne s'applique pas à l'épithète de l'espèce. (iii) Les noms des taxons doivent provenir du latin, du grec ou latinisés. (iv) Le nom d'un taxon sert en premier lieu à ce que l'on se réfère à lui, et non pas à indiquer les caractères ou l'histoire des taxons. (v) Chaque taxon possédant une délimitation, une position et un rang, ne peut porter qu'un seul nom. (vi) Le nom d'un taxon ne peut être changé que si des études taxonomiques l'indiquent comme erroné, ou si celui-ci contrevient aux règles du code.

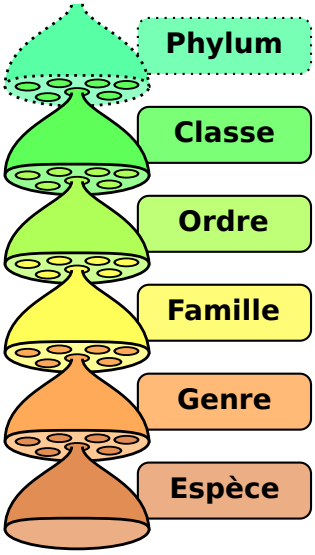


FIGURE A.VI – Rangs taxonomiques majeurs régulés par le code de nomenclature des procaryotes. Les phyla sont indiqués en pointillés, car non régis à l'heure actuelle par le code, bien qu'une proposition ait été effectuée (OREN et al. 2015).

D'après Peter Halasz, Wikipedia

Les règles du code régissent ensuite les noms des taxons des rangs taxonomiques indiqués dans la figure A.VI. L'ordre de ces rangs ne peut varier, contrairement à leur définition (par exemple, comme vu précédemment, la notion d'espèce bactérienne peut varier selon l'approche choisie). Il est important de noter que le nom des phyla ne sont pas concernés, au moment de l'écriture de ce manuscrit, par les règles du code, mais des propositions de changement des règles du code ont été formulées (OREN et al. 2015).

Les règles majeures du code sont présentées dans le tableau A.4. Celles-ci sont rétroactives (sauf exceptions) à chaque nouvelle révision. Le nom des

Rangs taxonomiques	Suffixe	Type nomenclatural
Sous-espèce		Souche type choisie lors de la première définition de l'espèce
Espèce		
Sous-genre		Espèce type choisie lors de la première définition du genre
Genre		
Sous-tribu	*-inae*	Genre type provenant du nom du taxon
Tribu	*-eae*	
Sous-famille	*-oideae*	
Famille	*-aceae*	
Sous-ordre	*-ineae*	
Ordre	*-ales*	
Sous-classe	*-idae*	
Classe	*-ia*	Choix d'un ordre type parmi ceux qui composent la classe
Phylum	*-aeota*	Choix d'une classe type parmi celles qui composent le phylum

Tableau A.4 – Récapitulatif des règles du code de nomenclature des procaryotes (Parker, Tindall et Garrity 2019). Les rangs taxonomiques en gras sont obligatoirement présents lors de la définition d'une espèce. La ligne en pointillés indique les propositions faites pour le rang taxonomique phylum.

espèces est binominal (Aubert 2016), c'est-à-dire qu'il est composé d'un nom générique combiné avec un épithète spécifique. Le nom générique doit être le nom du genre de l'espèce considérée (par exemple : *Escherichia*), tandis que l'épithète est libre (*coli*). La nomenclature binominale des espèces provient des travaux de Carl Linnaeus (Linné et Salvius 1758). Les autres rangs taxonomiques sont, eux, composés d'un mot. Le choix du nom d'un genre est libre, tant qu'il respecte les règles du code. Les noms des rangs supérieurs doivent être définis selon le genre type défini pour ceux-ci. Les noms de taxons inférieurs à la sous-espèce (tel que le nom des souches) ne sont pas régis par le code. C'est aussi le cas des termes associés aux taxons inférieurs à l'espèce (par exemple : pathovar, biovar, serovar, clone ou culture). Chaque rang taxonomique est associé à un taxon inférieur de manière permanente. On parle de type nomenclatural, et celui-ci doit être choisi lors de la définition du taxon. Dans le cadre de l'espèce, la souche type doit être cultivable si cela est possible pour l'une d'entre elles. Il existe également des souches de référence si la souche type d'une espèce n'est pas celle qui fait l'objet du plus grand

nombre d'études scientifiques. C'est le cas, par exemple, pour *Escherichia coli* dont la souche type est *Escherichia coli* DSM 30083 = JCM 1649 = ATCC 11775, tandis que *Escherichia coli* K12 MG1665 est une souche de référence.

Le code contient également les règles de publication et de validation d'un rang taxonomique. Afin qu'un nom soit valablement publié, il faut que celui-ci respecte les règles suivantes : (i) le nom doit être correct vis-à-vis du code, (ii) celui-ci doit être publié dans IJSEM, (iii) un type nomenclatural doit lui être assigné (Tableau A.4), (iv) le nom doit être accompagné d'une description, indiquant s'il s'agit de la description d'un nouveau taxon (sp. nov. : « *species nova* » pour une espèce) ou d'un changement de nom (comb. nov. : « *combinatio nova* »). La description doit également inclure l'étymologie si le nom est nouveau. Enfin, les propriétés intrinsèques du taxon doivent être décrites, (v) certains taxons possèdent des normes minimales afin de pouvoir décrire une nouvelle espèce [4]. (vi) depuis 2001, toute nouvelle description d'espèce est forcément accompagnée par le dépôt de la souche type dans deux collections de culture, dans deux pays différents. Ces collections doivent être accessibles librement. Pour les organismes présentant des risques sanitaires trop importants, des conditions de culture trop extrêmes ou incultivables, il est possible d'obtenir une dérogation au cas par cas.

Les principaux rôles des collections de culture sont de préserver et de distribuer les souches bactériennes et archéennes. Ce sont également des lieux de recherche sur la systématique et la taxonomie. Le tableau A.5 présente les abréviations de collections de cultures majeures. Celles-ci peuvent être associées avec un identifiant, afin de former un nom de souche permettant d'identifier dans quelles collections la souche a été déposée, comme par exemple *Escherichia coli* DSM 30083 = JCM 1649 = ATCC 11775 (Tableau A.5).

Le code reconnaît également le statut de « *Candidatus* », qui est un statut permettant de reconnaître les taxons pour lesquels une séquence d'acides aminés est disponible, ainsi que d'autres éléments attestant de son existence, mais qui ne peut remplir les critères du code car certains éléments sont manquants. Les critères nécessaires pour déclarer un taxon « *Candidatus* » sont disponibles dans l'appendice 11 du code de nomenclature et sont tirés des travaux de Murray et Schleifer (1994). Il est actuellement proposé que

4. À titre d'exemple, ces règles minimales pour la classe des *Mollicutes* imposent lors de la description d'une nouvelle espèce d'inclure (Brown, Whitcomb et Bradbury 2007 ; Whitcomb 2007) : (i) une description des interactions biologiques de l'organisme, (ii) une description de sa morphologie et de sa motilité, (iii) tester sa capacité à se développer à différentes températures et en présence ou non d'oxygène, (iv) effectuer des analyses sérologiques et (v) déposer l'antisérum de la souche type dans une collection. Ces derniers critères se rajoutent aux critères minimaux d'une description d'une espèce bactérienne.

lorsqu'un taxon est publié et validé, son nom candidat soit retenu en priorité sur les autres noms proposés lors du dépôt (Whitman, Sutcliffe et Rossello-Mora 2019).

Abréviation	Nom complet	Pays
ATCC	American Type Culture Collection	États-unis
CCUG	Culture Collection, University of Gothenburg	Suède
CIP	Collection de l'Institut Pasteur	France
DSM/DSMZ	Deutsche Sammlung von Mikroorganismenund Zellkulturen	Allemagne
JCM	Japanese Collection of Microorganisms	Japon
LMG/BCCM	Laboratory of Microbiology Ghent	Belgique
NCIMB	National Collections of Industrial, Marine and FoodBacteria	Royaume-uni
NCTC	National Collection of Type Cultures	Royaume-uni

Tableau A.5 – Sélection de collections de cultures majeures. L'ensemble des abréviations est disponible sur le site d'IJSEM.

D'après Lawson et al. (2016) et
`http://www.bacterio.net/-collections.html`

A.1.7 Les bases de données taxonomiques

Dans la partie précédente nous n'avons abordé que brièvement les règles du code. Bien que celles-ci soient fondamentales dans la microbiologie, elles souffrent d'une latence importante entre la découverte d'un nouveau taxon et sa publication valide. Les chercheurs en microbiologie font donc appel à des bases de données taxonomiques, afin de connaître la classification d'organismes d'intérêt. Il en existe actuellement huit majeures répertoriant des noms taxonomiques associés à une classification.

A.1.7.1 List of Prokaryotic names with Standing in Nomenclature

La List of Prokaryotic names with Standing in Nomenclature (LPSN, `http://www.bacterio.net/`, Euzéby 1997 ; Parte 2014 ; Parte 2018) est une liste incluant tous les noms publiés de manière valide dans IJSEM. Dans cette liste, chaque nom taxonomique est associé à la souche type, à un lien permettant d'accéder à la séquence du gène de l'ARNr 16S, ainsi que le lien

vers la publication ayant déclaré le nom. Le LPSN est une liste permettant de voir rapidement la validité d'un nom vis-à-vis de la nomenclature.

Aucune construction de classification n'est réalisé par le LPSN. La classification présentée sur le site provient du NCBI Taxonomy (Partie A.1.7.2, page 33), du « Taxonomic Outline of the Bacteria and Archaea », « Taxonomic Outlines » Volumes 3 et 4 du Bergey's Manual of Systematic Bacteriology (Second Edition) ainsi que du « The All-Species Living Tree Project » (Partie A.1.7.3, page 34).

A.1.7.2 NCBI Taxonomy

Le NCBI Taxonomy (`https://www.ncbi.nlm.nih.gov/taxonomy`, Federhen 2012) est la base de données taxonomiques de l'INSDC (formée par GenBank, Embl et DDBJ). Elle contient une nomenclature qui suit l'ICNP et une classification phylogénétique. La base de données a été créée en 1991 lors du projet « The Taxonomic Project » par le NCBI (Federhen 2003). Ce projet a permis de réunir plusieurs taxonomies et classifications phylogénétiques en une seule.

Cette base de données initiale a ensuite été et est toujours enrichie par le dépôt de nouvelles séquences dans l'une des trois banques de données de séquences de l'INSDC (GenBank, Embl et DDBJ). Si une nouvelle séquence déposée appartient à un nouveau taxon, alors son propriétaire doit prendre contact avec le comité du NCBI Taxonomy lors du dépôt, afin de justifier de sa position dans la classification. La littérature ayant trait à la systématique et à la phylogénie est la principale source de ce comité. Aucune analyse phylogénétique n'est menée par le NCBI Taxonomy. Chaque nom de taxon est lié à un identifiant (TaxID) stable dans le temps. En 2014, la décision a été prise que les souches ne se verraient plus associées à un TaxID, car il n'était plus possible d'assurer leur curation (Federhen et al. 2014). Chaque taxon est lié à son type nomenclatural, à son nom formel, à ses noms informels, aux séquences se rapportant à ce taxon, à la littérature propre à ce taxon, ainsi qu'à des liens externes permettant d'explorer les informations disponibles sur ce taxon dans d'autres banques de données, via le service LinkOut.

Le NCBI Taxonomy positionne également dans sa classification des organismes non cultivés et non publiés dans IJSEM. C'est le cas notamment des organismes associés aux séquences obtenues lors de prélèvements environnementaux. Ces organismes présentent donc une classification hypothétique. Le nombre de noms de taxon est présenté dans le tableau A.6.

La figure A.VII montre l'évolution du nombre de noms de genres et d'espèces publiés et validés dans le NCBI Taxonomy.

	Genre	Espèce
Archaea	176	671
Bacteria	3 426	18 354

Tableau A.6 – Nombre de noms d'espèces et de genres dans le NCBI Taxonomy. Ces chiffres excluent les noms d'espèces et de genres non-classifiés, non cultivés et les noms informels.

Source: NCBI Taxonomy. Consulté le 30/08/2019

A.1.7.3 The All-Species Living Tree Project

The All-Species Living Tree Project (LTP, `https://www.arb-silva.de/projects/living-tree/`) a débuté en 2007 et son but est de fournir une taxonomie phylogénétique basée sur le gène de l'ARNr 16S (Yarza et al. 2008 ; Yarza et al. 2010 ; Munoz et al. 2011). Il réunit des données provenant de plusieurs entités : (i) le LPSN et le *Bergey's Manual of Systematic Bacteriology* servent de sources pour les noms de taxons, (ii) les journaux *Systematic and Applied Microbiology* et IJSEM, et (iii) les séquences complètes du gène de l'ARNr proviennent de la base SILVA (Quast et al. 2013). Une seule séquence est considérée, sauf si les différentes copies sont divergentes de plus de 98%. Les séquences des souches types sont associées à leur classification du LPSN. Cet ensemble de séquences est ensuite utilisé afin de générer un arbre phylogénétique, grâce à la méthode du maximum de vraisemblance et au logiciel RAxML (Stamatakis 2014). Cet arbre est entièrement recalculé tous les deux ans environ. L'arbre est ensuite examiné afin d'évaluer la monophylie de chaque taxon. Les clades se voient ensuite attribuer un nom selon les genres types présents dans celui-ci. Cet arbre est, dans sa version 132, composé de 13 903 séquences d'ARNr 16S de souches types. Cet arbre est de petite taille, mais comme celui-ci est formé à partir de l'intégralité des séquences d'ARNr 16S de souches types, il est considéré comme exhaustif taxonomiquement en ce qui concerne les souches types .

A.1.7.4 La taxonomie SILVA

La taxonomie SILVA (`https://www.arb-silva.de/documentation/silva-taxonomy/`) est basée sur un arbre phylogénétique guide, de séquences du gène de l'ARNr 16S initié en 2004 par le projet ARB (Ludwig et al. 2004 ; Yilmaz et al. 2014), auquel sont rajoutés régulièrement des nouvelles séquences. La curation du placement des nouvelles séquences est réalisée *a posteriori*, en incluant les retours de la communauté d'utilisateurs. Les clades

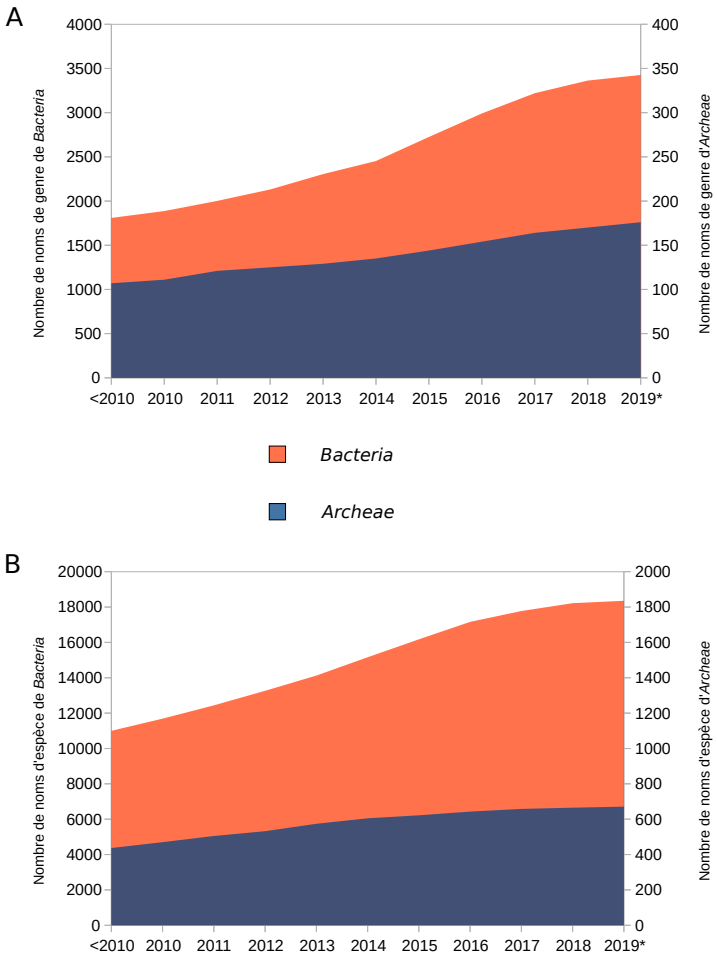


FIGURE A.VII – Évolution du nombre de noms de genres (**A**) et d'espèces (**B**) dans le NCBI Taxonomy. * : Le nombre pour 2019 a été arrêté au 30/08/2019.

sont affiliés à des noms de taxons issus des *Bergey's Taxonomic Outlines* (GARRITY, BELL et LILBURN 2004), qui est un index des taxons contenus dans le *Bergey's Manual of Systematic Bacteriology*. Le LPSN est également utilisé afin d'obtenir les noms de taxons actuels et conformes à la littérature actuelle, ainsi que pour corriger les noms non valides et « *Candidatus* ». Les *Bergey's Taxonomic Outlines* et le LPSN n'étant pas uniquement basés sur un arbre phylogénétique mais sur une classification polyphasique, il peut exister des contradictions entre ces derniers et la taxonomie SILVA. La taxonomie SILVA fait également appel aux avancées de classification des arbres phylogénétiques basés sur des protéines marqueurs (GTDB, Partie A.1.7.8, page 37) qui ont mené à des changements drastiques dans le phylum des

proteobactéries (Parks et al. 2018), ainsi qu'à des sources spécialisées dans certains clades, afin de préciser certaines parties de sa classification et afin de palier le manque de pouvoir discriminant du gène de l'ARNr 16S pour certains taxons.

A.1.7.5 Greengenes

Greengenes (`https://greengenes.secondgenome.com`) est une base de données de séquences de gènes de l'ARNr 16S qui contient également une classification et une taxonomie (McDonald et al. 2012). Un arbre phylogénétique de l'intégralité des séquences 16S contenues dans Greengenes est calculé et un script permet d'assigner aux clades un rang taxonomique provenant du NCBI Taxonomy. Cette base de données n'est plus mise à jour depuis mai 2013.

A.1.7.6 The Ribosomal Database Project

The Ribosomal Database Project (RDP, `https://rdp.cme.msu.edu/`) est à la fois une base de données de séquences alignées et annotées de gènes de l'ARNr 16S et une compilation d'outils permettant d'analyser ces séquences ou des séquences extérieures (Cole et al. 2014). La taxonomie RDP est basée sur sur le *Bergey's Taxonomic Outline of the Prokaryotes* (Garrity, Bell et Lilburn 2004). L'ensemble des gènes de l'ARNr 16S des espèces types sont tirés de GenBank et sont alignés les uns contre les autres. RDP classifier (Wang et al. 2007) permet ensuite d'assigner à de nouvelles séquences une hiérarchie. La taxonomie RDP ne contient pas le rang d'espèce et assigne les séquences au genre le plus proche.

A.1.7.7 Open Tree of Life

Open Tree of Life (OTL, `https://tree.opentreeoflife.org/`) est un projet de construction de super arbre phylogénétique représentant l'intégralité des organismes vivants (Hinchliff et al. 2015). Celui-ci est composé d'arbres phylogénétiques de plus petite taille obtenues depuis Tree Base (Piel, Donoghue et Sanderson 2000), Dryad (site de dépôt de données scientifiques) et directement auprès d'auteurs d'arbres phylogénétiques. En parallèle de ce super arbre est construit l'OpenTree Taxonomy (OTT) (Rees et Cranston 2017) qui est composé de plusieurs sources de taxonomie, notamment pour les procaryotes : SILVA, le NCBI Taxonomy, Global Biodiversity Information Facility (GBIF) et l'Interim Register of Marine and Nonmarine Genera (IRMNG). Ces taxonomies sont rassemblées et converties en un format commun. Chaque noeud se voit attribuer un identifiant unique.

L'arbre phylogénétique et la taxonomie sont ensuite rassemblés pour former un arbre de la vie.

A.1.7.8 Genome Taxonomy DataBase

Genome Taxonomy DataBase (GTDB, `https://gtdb.ecogenomic.org/`) est une base de données taxonomique (Parks et al. 2018). Un arbre phylogénétique est inféré en utilisant l'intégralité des génomes bactériens et d'archées provenant de RefSeq et en s'appuyant sur 120 protéines bactériennes marqueurs (122 protéines pour les archées). Le NCBI Taxonomy est ensuite utilisé afin de donner un nom aux clades de l'arbre. Certaines régions sans organismes cultivés nécessitent l'utilisation de GreenGenes ou SILVA, afin de pouvoir nommer ces régions. Le LPSN est utilisé pour la curation des synonymes. Cette base de données est indexée sur les nouvelles versions de RefSeq. A chaque nouvelle version, la base de données est recalculée. A la suite du calcul de cet arbre, 58% des génomes provenant de RefSeq ont subi un changement de taxonomie dans GTDB qui au moment de l'écriture de ce manuscrit n'a pas été répercuté sur le NCBI Taxonomy.

	Origine des noms taxonomiques	**Méthode de construction de la classification**	**Points forts**	**Points faibles**
LPSN	Littérature	Aucune	Au plus proche de la littérature	Mise à jour lente
NCBI Tax.	Genbank	Aucune, curation lors du dépôt d'un nouveau taxon	Intéropérable et lié à une banque de données génomique	Pas d'analyse phylogénétique direct
LTP	LPSN et *Bergey's Taxonomic Outline of the Prokaryotes*	Inférence phylogénétique à partir des séquences 16S de souche type	Souche type	Manque de diversité
SILVA Tax.	LPSN et *Bergey's Taxonomic Outline of the Prokaryotes*	Inférence phylogénétique à partir des séquences 16S et d'un arbre guide	Classification mise à jour grâce à de nouvelles données	Spécialisé dans le gène de l'ARNr 16S
Greengenes	NCBI Taxonomy	Inférence phylogénétique à partir des séquences 16S		Non mise à jour
RDP Tax.	*Bergey's Taxonomic Outline of the Prokaryotes*	Alignement de séquence du gène de l'ARNr 16S	Classification reproductible	Basé sur l'ARNr 16S
OTL	NCBI Taxonomy, SILVA, GBIF, IRMNG, ...	Inférence d'un super arbre phylogénétique	Diversité	Pas de curation manuelle et non relié à des séquences
GTDB	NCBI Taxonomy	Inférence d'un arbre phylogénétique de génomes bactériens grâce aux séquences protéiques de gènes marqueurs	Utilisation de multiples séquences protéiques de gènes marqueurs	

Tableau A.7 – Comparaison des méthodes de construction de la classification dans les huit bases de données taxonomiques majeures, ainsi que l'origine des noms taxonomiques.

A.1.7.9 Analyse comparée des bases de données taxonomiques

Les différentes méthodes de construction des classifications, ainsi que la source des noms taxonomiques sont résumées dans le tableau A.7.

Dans Balvočiūtė et Huson (2017), cinq classifications taxonomiques ont été analysées. Les auteurs ont identifié les noms de genres communs en totalité ou partiellement ou uniques aux cinq classifications.

La figure A.VIII présente les résultats de cette correspondance entre taxonomies. On observe que le NCBI Taxonomy partage plus de noms taxonomiques avec SILVA qu'avec RDP et Greengenes. Mais peu de noms sont partagés entre toutes les taxonomies : 73% des phyla, 70% des classes, 63% des ordres, 90% des familles et 89% des genres ne sont pas partagés et sont uniques soit à SILVA, RDP, Greengenes ou au NCBI Taxonomy (OTT est exclu de ce calcul). OTT contient quant à elle plus de noms que l'union des quatre bases taxonomiques au niveau du genre.

La figure A.IX, page 41, présente les indices de dissimilarité entre les différentes bases de données selon que l'on applique une correspondance stricte (« Strict mapping ») ou une correspondance lâche (« Loose mapping »). En correspondance stricte Greengenes présente un indice de dissimilarité le plus faible et proche de 0.25. La seule autre valeur proche de 0.25 est la correspondance du NCBI Taxonomy sur OTT. Cela peut s'expliquer par le fait que OTT est formé en grande partie par le NCBI Taxonomy. En correspondance lâche on peut observer que le NCBI Taxonomy et OTT présentent une forte similitude quel que soit le sens de correspondance.

Les principales conclusions de cet article sont que les taxonomies SILVA, Greengenes et RDP peuvent facilement correspondre avec celles du NCBI Taxonomy et ces quatres bases de données correspondent facilement avec les données d'OTT. L'inverse (faire correspondre le NCBI Taxonomy sur OTT par exemple) est problématique. Un des désavantages d'OTT est qu'il n'est pas associé avec une banque de données de séquences, ce qui rend difficile son utilisation pour l'affiliation de lecture de séquençage.

Aucune de ces bases n'est donc parfaite, et de nombreuses erreurs dans la conception et leur taxonomie ont été relevées, en particulier dans les bases de données reposant sur le gène de l'ARNr 16S (Edgar 2018a). L'arbre guide de SILVA présente par exemple des incohérences avec l'arbre Greengenes. Environ 17% des annotations dans Greengenes et dans SILVA seraient en conflit. Les classifications sont peu reproductibles hormis celle de RDP. Parmi toutes les bases de données taxonomiques ici présentées, le NCBI Taxonomy se révèle comme étant centrale par rapport aux autres car cette base de données permet de relier les séquences génomiques entre elles (ARNr 16S, autres gènes voire génome complet) grâce à un identifiant unique.

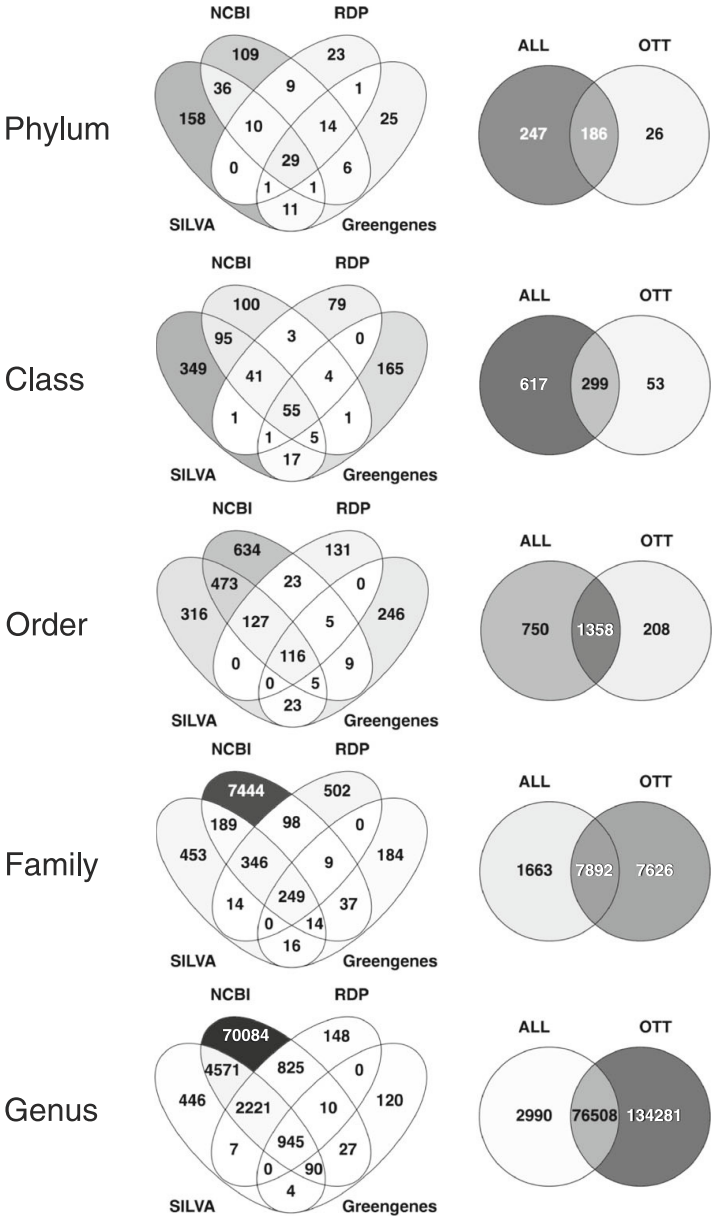


FIGURE A.VIII – Diagrammes de Venn de l'intégralité des noms de taxons (formels et informels) présents dans une ou plusieurs bases de données taxonomiques. Les diagrammes de gauche présentent, pour chaque rang taxonomique, les noms présents dans NCBI Taxonomy, Greengenes, SILVA et RDP. Le diagramme de droite présente l'union des taxonomies précédentes contre OTT. Les nuances de gris indiquent le pourcentage de noms taxonomiques contenus dans l'intersection.

Source: BALVOČIŪTĖ et HUSON (2017)

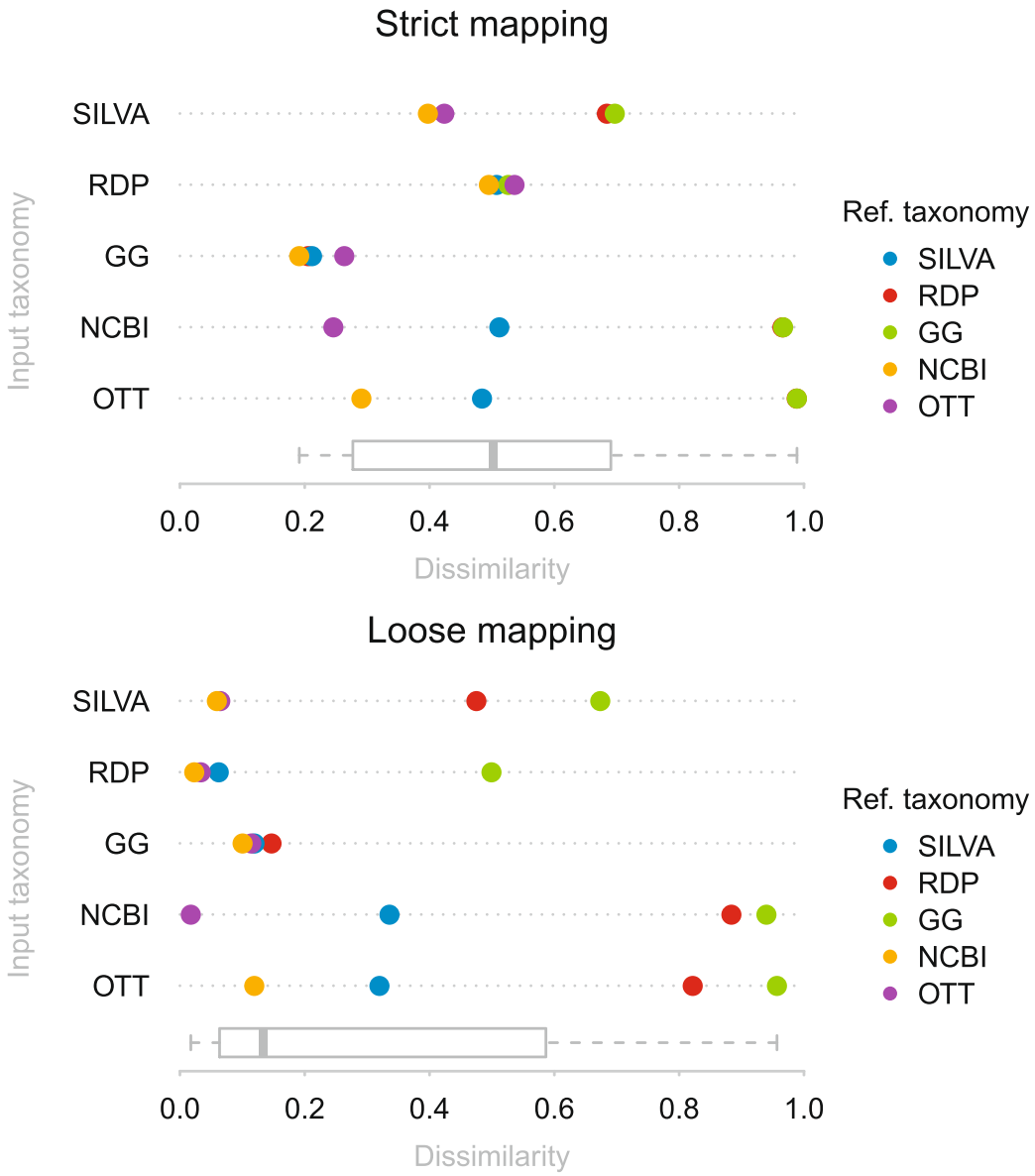


FIGURE A.IX – Dissimilarité entre les cinq bases de données taxonomiques fondées sur les correspondances de noms de taxons par paires. Plus l'indice de dissimilarité est proche de 1, plus les taxonomies sont différentes. « Strict mapping » : un même nom taxonomique doit se trouver dans les deux bases au même rang. Dès qu'un nom taxonomique est absent, les rangs taxonomiques inférieurs sont également considérés comme incorrectes. « Loose mapping » : reprend le principe précédent mais si un nom taxonomique est absent, les rangs fils sont considérés indépendamment.

D'après BALVOČIŪTĖ et HUSON (2017)

A.2 Méthodes de détermination de la composition procaryotique d'un environnement

La culture de souche pure était, dans le passé, la seule méthode d'identification des procaryotes. La culture microbiologique était également utilisée pour identifier les espèces et les quantifier dans un environnement complexe. Cette stratégie a toutefois montré ses limites dans les années 1990, les descriptions d'espèces bactériennes cultivables devenant de plus en plus rares. En effet, toutes les bactéries ne sont pas cultivables dans les milieux de culture « classiques »car certaines demandent des nutriments et/ou des conditions environnementales particulières. Les bactéries forment également des communautés, et une espèce bactérienne peut dépendre d'une autre pour sa survie (Vartoukian, Palmer et Wade 2010 ; Stewart 2012). Dans le cas de culture d'échantillons environnementaux, la culture microbiologique possède un biais qui ne permet pas d'évaluer la quantité de toutes les espèces présentes (The Great Plate Count Anomaly, Amann, Ludwig et Schleifer 1995). En 1985, bien que l'exploration des espèces bactériennes incultivables soit encore un problème mineur, les premières expérimentations d'exploration de la diversité bactérienne ont lieu, grâce aux avancées en biologie moléculaire, parmi lesquelles le séquençage de gènes amplicons, notamment du gène de l'ARNr 16S (Partie A.1.5.2, page 17) (Olsen et al. 1986 ; Pace et al. 1986). Avec l'avènement du séquençage à haut débit, les communautés bactériennes peuvent désormais être analysées par séquençage de l'intégralité de l'ADN (Partie A.1.5.2, page 24) présent dans une communauté et par l'assemblage des génomes des organismes présents (Venter et al. 2004 ; Tyson et al. 2004). Aujourd'hui, ces deux approches sont utilisées de manière routinière pour déterminer la composition des communautés bactériennes. Nous présentons ici ces deux approches d'affiliation taxonomique de procaryotes au sein d'un écosystème : par amplicons et par séquençage de l'intégralité de l'ADN présent dans un écosystème.

A.2.1 Approche amplicon

L'approche amplicon nécessite l'amplification par PCR du gène marqueur d'intérêt (pour l'identification de la composition d'une communauté procaryotique, il s'agit couramment de gène de l'ARNr 16S) et son séquençage. Les lectures sont ensuite réunies dans des groupes de séquences proches appelés Unités Taxonomiques Opérationnelles (OTU, de l'anglais : *Operational Taxonomic Unit*).

A.2.1.1 OTU

Le terme OTU a été utilisé pour la première fois par Sokal (1963) dans le cadre de la taxonomie numérique (Partie A.1.4, page 13). Ce terme évoquait alors un ensemble d'organismes présentant des similarités. Dans le cadre d'analyses de communautés procaryotiques par gène amplicon, une OTU est l'ensemble des séquences proches, c'est-à-dire ayant une suite de nucléotides similaires. Ce concept présente plusieurs avantages : (i) réunir les séquences en groupes et ne prendre qu'une seule séquence représentative pour chaque groupe permet de gagner en temps de calcul pour l'affiliation taxonomique. La taille du jeu de données passe de millions de lectures en sortie de séquençage à quelques milliers. (ii) Cela permet de se rapprocher de la définition de l'espèce bactérienne, même si l'OTU ne peut prétendre à ce statut. (iii) Permet de regrouper les séquences altérées lors des biais techniques (PCR et séquençage par exemple) à la séquence graine.

A.2.1.2 Techniques de groupement de séquence

Par pourcentage de similarité de séquence La première technique de groupement de séquence (de l'anglais : *clustering*) repose sur la similarité des séquences entre elles. Des groupes (clusters) sont formés par alignement de séquences. La similarité entre deux séquences est calculée en pourcentage. Le seuil de similarité par défaut des outils, communément accepté par la communauté scientifique, est de 97% de similarité de séquence, pour les séquences du gène de l'ARNr 16S. Ce seuil est dérivé du seuil de l'espèce bactérienne précédemment évoqué (Partie A.1.5.2, page 17), ainsi que d'une étude empirique sur la similarité de séquence du gène de l'ARN 16S entre souches (Konstantinidis et Tiedje 2005a). De nombreux outils permettant le groupement de séquences existent, parmi lesquels : VSEARCH (Rognes et al. 2016), UPARSE (Edgar 2013), mothur (Schloss et al. 2009) et QIIME 2 par l'intermédiaire de ses extensions (Bolyen et al. 2019). Cette technique fait actuellement l'objet de critiques (Nguyen et al. 2016 ; Edgar 2018b), dont notamment le seuil de 97% (Mysara et al. 2017) ainsi que la dépendance à la séquence initiale utilisée pour former chaque groupe.

Il existe des méthodes sans alignement reposant sur les *k-mers*, mots de petite taille fixe. Ceux-ci peuvent être indexés très rapidement et être utilisés comme unités de comparaison entre deux séquences. RDP et KRAKEN 2 (Wood et Salzberg 2014) utilisent cette méthode.

Par essaim Il est également possible de former des clusters par la méthode de l'essaim dans des outils tels que Swarm (Mahé et al. 2015) ou

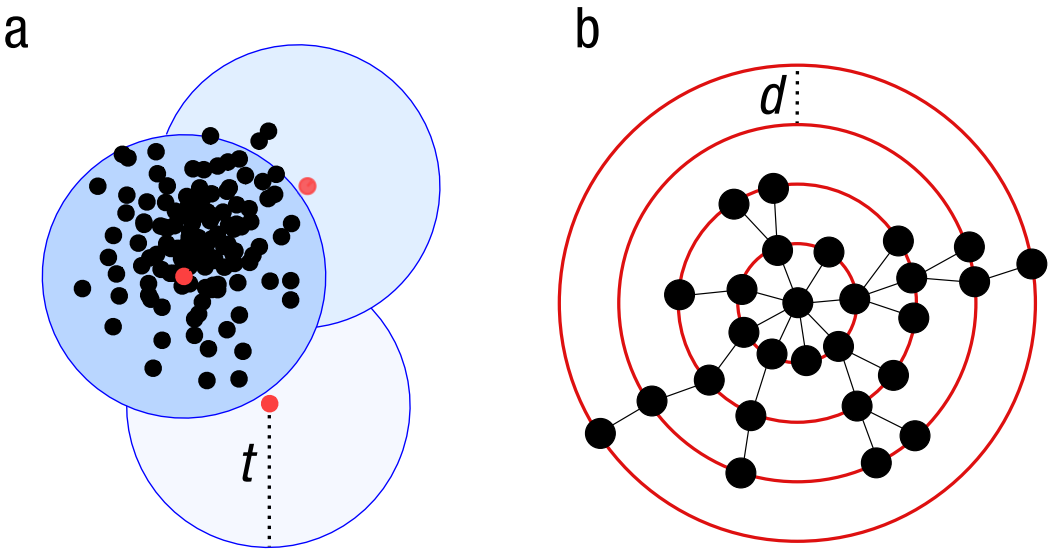


FIGURE A.X – Schéma de deux méthodes de groupement de séquences. A : par identité de séquence. Les ronds rouges sont les séquences centrales choisies au hasard pour former un groupe à $\boldsymbol{t}$ % de similarité. Bien que les différentes séquences soient proches, elles appartiennent à différents groupes. B : par essaim. Une séquence est choisie au hasard et toutes les séquences avec $\boldsymbol{d}$ nucléotides de différence sont ajoutées aux groupes. Une fois cette opération réalisée, elle est répétée pour les séquences ajoutées. Toutes les séquences appartiennent ici à la même OTU.

Source: MAHÉ et al. (2014)

SUMACLUST (MERCIER et al. 2013). Elle se base sur un nombre maximal de différences entre séquences, et non sur un pourcentage de similarité (Figure A.X). Bien que cette méthode s'éloigne de la définition de l'espèce bactérienne, elle est tout à fait adaptée aux problématiques d'erreurs de séquencage introduites dans les séquences cycle après cycle d'amplification de PCR. De plus, ces algorithmes sont indépendants de l'ordre d'entrée des séquences à clusteriser, contrairement aux algorithmes qui agglomèrent les séquences par leur similarité (Partie A.2.1.2, page 43).

Amplicon Sequence Variants La méthode des « Amplicon Sequence Variants » (ASV) est une alternative à la construction d'OTUs utilisée dans des logiciels tels que DADA2 (CALLAHAN et al. 2016) et Deblur (AMIR et al. 2017). Celle-ci consiste à corriger les lectures de leurs erreurs de séquençage et de PCR (biais techniques connus). Ces séquences corrigées permettent d'inférer des séquences uniques présentes dans l'échantillon original. Cette méthode

a été décrite comme permettant d'atteindre une résolution de séquence plus fine ainsi qu'une plus grande reproductibilité (Callahan, McMurdie et Holmes 2017).

A.2.1.3 Affiliation taxonomique

Une fois les groupes de séquences formées, leurs séquences représentatives sont taxonomiquement affiliées. Pour cela, le choix d'une base de données de séquences d'amplicons associée à une taxonomie est nécessaire. On peut citer par exemple SILVA, Greengenes, RDP, ou encore GenBank (Partie A.1.7, page 32). Un outil d'affiliation est également nécessaire, tel que BLAST+ (Camacho et al. 2009), RDP classifier (Wang et al. 2007) ou encore classify-sklearn de QIIME 2 (Bolyen et al. 2019). Certains outils prennent en charge l'intégralité des étapes des lectures brutes jusqu'à l'obtention de la table d'abondance représentant l'abondance relative des groupes associés avec leur affiliation taxonomique. On citera notamment FROGS (Escudié et al. 2018), UPARSE (Edgar 2013), Mothur (Schloss et al. 2009), DADA2 (Callahan et al. 2016) et QIIME 2 (Bolyen et al. 2019).

A.2.1.4 Avantages et inconvénients

Les avantages de l'approche amplicon sont multiples : (i) le coût du séquençage est peu élevé et permet d'obtenir une profondeur de séquençage élevée, (ii) de nombreux outils existent permettant une prise en charge rapide des lectures issues du séquençage. Ces outils demandent peu de moyens informatiques. (iii) Des banques de données contenant un nombre important de séquences du gène d'ARNr 16S existent et permettent d'affilier facilement les séquences. Mais cette méthode souffre aussi de certaines limites : (i) le choix du gène marqueur dépend de la composition de la communauté bactérienne. Le choix de la région variable, en ce qui concerne le 16S, est également primordial (Edgar 2018b ; Bukin et al. 2019), (ii) la résolution est limitée à l'espèce mais à ce niveau elle est considérée comme peu fiable, (iii) le nombre de copies du gène de l'ARNr 16S est variable selon les organismes, (iv) accès à une abondance relative des différents taxons et non à une abondance absolue. Ceci est dû au biais d'amplification lors de la PCR.

A.2.2 Approche par séquençage métagénomique

L'avancée des techniques de séquençage ont rapidement permis de séquencer l'intégralité de l'ADN présent dans une communauté bactérienne et un environnement donnés. Cette technique génère des millions de lectures qu'il

convient ensuite de traiter afin de déterminer la composition microbienne du milieu. La métagénomique permet de mettre en évidence le contenu en gènes d'une population microbienne et de déterminer ainsi leurs capacités fonctionnelles. Cette approche demande plus de ressources de séquençage et est donc moins sensible que la précédente pour étudier la biodiversité. Deux approches majeures existent afin d'analyser ces lectures.

A.2.2.1 Approche avec référence

Cette approche implique d'aligner les lectures de séquençage contre des séquences de références affiliées taxonomiquement, le plus souvent des génomes complets d'organismes, via notamment BLAST (Altschul et al. 1990). Il est également possible d'assembler *de novo* les lectures en contigs. Cette étape est optionnelle mais permet d'obtenir des séquences de plus grande taille facilitant l'affiliation taxonomique. Cette étape fait appel à des outils spécialisés dans l'assemblage de lectures provenant de métagénomes dont MegaHit (Li et al. 2016) et metaSPAdes (Nurk et al. 2017) entre autres (Ayling, Clark et Leggett 2019).

Cette approche est très performantes pour reconnaître les espèces présentes connues, mais reste extrêmement limitée dans la découverte de nouvelles espèces, les séquences de références n'existant pas.

A.2.2.2 Approche sans référence

Il existent deux approches sans référence majeure. La première nécessite un assemblage des lectures en contigs. Ces contigs peuvent ensuite être assemblés en génomes grâce à la cooccurence de l'abondance des contigs dans le métagénome. On parle alors de génomes assemblés par métagénomique (MAG, de l'anglais : *Metagenome-Assembled Genomes*). Mais l'assemblage de MAGs reste une technique en cours de développement, les espèces rares n'étant pas assemblées et la diversité des souches ne pouvant être déterminé (Sczyrba et al. 2017).

Le *binning* ou groupement de lectures métagénomiques sans référence est une technique rappelant le groupement des séquences amplicons en OTUs. Elle consiste à rassembler les lectures métagénomiques (ou les contigs si les lectures sont assemblées dans un premier temps) dans des groupes appelés *bin*. Il existe pour cela plusieurs algorithmes (Breitwieser, Lu et Salzberg 2017). Chaque *bin* est censée être constituée d'un seul taxon et peut alors être assemblée et/ou affiliée taxonomiquement. Si la *bin* contient des génomes reconstitués après assemblage, il est possible de parler d'espèces métagénomiques (MGS, de l'anglais : *MetaGenomic Species*, Zepeda

Mendoza, Sicheritz-Pontén et Gilbert 2015).

Les méthodes sans référence permettent d'extraire des organismes taxonomiquement rares ou inconnus. Ceci ouvre la porte à une meilleure compréhension des communautés bactériennes.

A.2.2.3 Avantages et inconvénients

L'approche métagénomique via séquençage métagénomique propose plusieurs avantages par rapport à l'approche amplicon : (i) plus de biais d'amplification. Les abondances des lectures sont quantitatives au vu du nombre de lectures séquencées. (ii) Elle permet de s'approcher de la reconstruction intégrale des génomes des organismes présents dans l'écosystème et d'effectuer des analyses plus approfondies des génomes reconstruits sur celui-ci notamment au niveau des gènes. (iii) La résolution taxonomique est au niveau de l'espèce, voire de la souche dans certains cas. Mais cette approche présente aussi des limites : (i) un coût de séquençage supérieur avec l'obtention d'une profondeur de séquençage plus faible que l'approche amplicon, (ii) l'analyse des lectures est longue et demande une puissance de calcul suffisante (Vincent et al. 2017), (iii) les génomes complets de référence sont présents en faible nombre dans les bases de données ce qui freine l'affiliation taxonomique.

A.3 Exploration du potentiel fonctionnel des procaryotes

A.3.1 Le métabolisme procaryote

Le métabolisme (du grec $\mu\epsilon\tau\alpha\beta o\lambda\acute{\eta}$, metabolê : changement) désigne l'ensemble des réactions biochimiques se produisant dans une cellule. Ces réactions biochimiques permettent de transformer des composés chimiques de petites tailles moléculaires (< 1 500 Da) - appelés métabolites - ou des macromolécules en un ou plusieurs autres métabolites/macromolécules. Ces réactions peuvent également produire ou nécessiter de l'énergie et sont, dans une majorité des cas, catalysées par des enzymes qui leur sont spécifiques. Quand plusieurs réactions sont liées en série entre elles, elles forment une voie métabolique (Figure A.XI) telles que : (i) le cycle de Krebs permettant de dégrader des glucides, des graisses et des protéines afin d'obtenir de l'énergie chez les organismes anaérobiques, (ii) la voie d'Embden-Meyerhof-Parnas, aussi appelée glycolyse, produisant de l'énergie en assimilant le glucose, (iii) la fermentation de la L-lysine en acétate et butyrate.

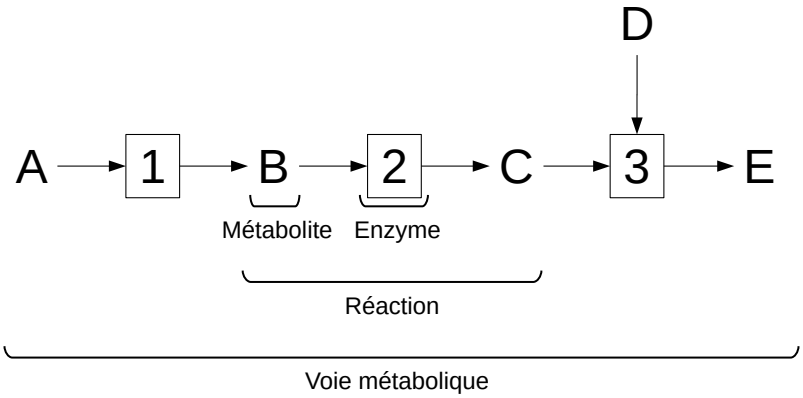


FIGURE A.XI – Les acteurs du métabolisme. Les lettres représentent des métabolites tandis que les chiffres représentent des enzymes. L'ensemble métabolite(s) d'entrée (substrat(s), **B**), enzyme (**2**) et métabolite(s) de sortie (produit(s), **C**) est appelé réaction. Une suite de réactions forme une voie métabolique.

Le métabolisme d'une cellule est divisé en deux types : (i) le métabolisme primaire (HODGSON 2000), impliqué dans la croissance, le développement et la reproduction de la cellule. Ce métabolisme contient l'ensemble des voies (nommées voies métaboliques primaires) et des réactions nécessaires à la vie. Bien que le métabolisme primaire soit étudié depuis plus d'un siècle,

certaines parties sont encore inconnues (Nielsen 2017). (ii) Le métabolisme secondaire (aussi appelé métabolisme spécialisé) (Haslam 1986) contient les voies métaboliques (secondaires) qui ne sont pas indispensables à la vie de la cellule ou de l'organisme, comme par exemple la production d'antibiotiques. La difficulté de l'étude du métabolisme secondaire provient du fait qu'il est spécifique à des groupes taxonomiques précis et présente des métabolites chimiquement très diverses (Bérdy 2005).

Le métabolisme peut également se décomposer en deux parties : (i) une partie catabolique, constituée de voies métaboliques dégradant des macromolécules en énergie (ATP, NADH) et/ou en métabolites plus petits pouvant être utilisés dans des réactions de biosynthèse. On peut citer comme voies majeures du catabolisme la voie d'Embden-Meyerhof-Parnas, dégradant le glucose afin de produire de l'énergie et du pyruvate, ou encore le cycle de Krebs produisant de l'énergie et des précurseurs aux acides aminés non-essentiels. (ii) Une partie anabolique, où des voies métaboliques de biosynthèse utilisent les métabolites précurseurs produits lors du catabolisme afin de synthétiser des composants cellulaires tels que les protéines, les acides nucléiques ou les polysaccharides. La biosynthèse de ces éléments entraîne la consommation d'énergie produite lors de la partie catabolique du métabolisme. On peut citer, par exemple, l'ensemble des voies permettant la synthèse des acides aminés ou la voie de la biosynthèse du peptidoglycane chez les bactéries.

Les bactéries ont une diversité métabolique plus grande que les eucaryotes (Oren 2009) et l'Homme utilise cette diversité métabolique procaryote à son avantage, par exemple en exploitant certains genres bactériens possédant des voies métaboliques impactant le goût du fromage, tels que les genres *Lactobacillus*, *Streptococcus* ou encore *Lactococcus* (Smit, Smit et Engels 2005 ; Hanniffy et al. 2009). Les archées n'ont pas, à l'heure actuelle, d'application directe dans l'industrie malgré des perspectives encourageantes (Schiraldi, Giuliano et De Rosa 2002 ; Cabrera et Blamey 2018).

A.3.2 Les acteurs du métabolisme

Dans cette partie nous nous attarderons sur les différents acteurs composant le métabolisme.

A.3.2.1 Les métabolites

Les métabolites sont présents dans les différents compartiments de la cellule. Ils sont les intermédiaires et les produits des réactions chimiques. Un métabolite peut aussi être le substrat (élément de départ) d'une réaction métabolique, mais les réactants peuvent aussi être importés de l'extérieur de

la cellule. On parle alors de nutriments (dans le cas de molécules apportées par l'alimentation) ou de xénobiotiques (dans le cas de composés étrangers à la cellule).

Métabolite	**Abréviation**	**Briques élémentaires et macromolécules produites**
D-glucose-6-phosphate	G6P	Glycogène, lipopolysaccharides
D-fructose-6-phosphate	F6P	Paroi celullaire (peptidoglycane)
D-ribose-5-phosphate	R5P	His, Phe, Trp, nucléotides
D-erythrose-4-phosphate	E4P	Phe, Trp, Tyr
D-glyceraldehyde-3-phosphate	GAP	Lipides
glycerate-3-phosphate	3PG	Cys, Gly, Ser
phosphoenolpyruvate	PEP	Tyr, Trp
pyruvate	PYR	Ala, Ile, Lys, Leu, Val
acetyl-CoA	ACA	Leu, lipides
2-ketoglutarate	2KG	Glu, Gln, Arg, Pro
succinyl-CoA	SCA	Met, Lys, tétrapyrrole (par ex., l'hème)
oxaloacetate	OXA	Asn, Asp, Ile, Lys, Met, Thr, nucléotides

Tableau A.8 – Les douze métabolites précurseurs ainsi que leur devenir dans la production de briques élémentaires et macromolécules chez *Escherichia coli*.

Source: Adapté de Noor et al. (2010)

Le métabolisme se base sur 12 métabolites précurseurs (Tableau A.8, Noor et al. 2010) qui sont issus de la dégradation du glucose (via les voies de la glycolyse, des pentoses phosphates et du cycle de Krebs entre autres) pour les organismes hétérotrophes ou via la fixation du CO^2 pour les organismes autotrophes. Ces précurseurs sont les squelettes carbonés nécessaires pour la synthèse de tous les monomères et « briques élémentaires » présents dans la cellule.

Les « briques élémentaires » sont des métabolites qui en se polymérisant forment des macromolécules essentielles à la vie. Il en existe une cinquantaine, dont les acides aminés, les acides gras et les nucléotides (Nielsen 2017).

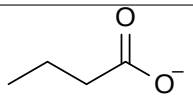
Nom d'usage	butyrate
Nom IUPAC	butanoate
Formule	$C_4H_7O_2-$
Numéro CAS	107-92-6
Structure	
InChI	1S/C4H8O2/c1-2-3-4(5)6/h2-3H2,1H3,(H,5,6)/p-1

Tableau A.9 – Nom IUPAC, formule, numéro CAS, structure et IUPAC International Chemical Identifier du butyrate.

Chaque métabolite est défini par sa formule et sa structure chimique. Chaque métabolite est lié à un numéro de registre CAS (Chemical Abstract Service), qui est un identifiant unique associé à des composés chimiques (pas uniquement aux métabolites). Les noms des métabolites devraient, quant à eux, suivre la nomenclature de l'IUPAC (International Union of Pure and Applied Chemistry, Leigh 2011), mais cette nomenclature donnant des noms complexes pour les molécules, est rarement utilisée par les biologistes. L'identifiant InChI (IUPAC International Chemical Identifier Heller et al. 2013 permet de connaître la formule d'un métabolite (ou de toute autre substance chimique), ainsi que sa structure. Le tableau A.9 contient, comme exemple l'ensemble de ces informations pour le butyrate, un acide gras à chaîne courte produit par les bactéries du microbiote intestinal (Morrison et Preston 2016).

L'ensemble des métabolites d'un organisme est appelé le métabolome. Les métabolites y sont classés en tant que métabolites primaires ou métabolites secondaires (ayant le même sens que le métabolisme primaire et secondaire). A titre d'exemple, 3 755 métabolites ont été décrits chez *Escherichia coli* K-12 MG1655 au moment de l'écriture de ce manuscrit (Sajed et al. 2016).

L'étude du métabolome d'un organisme se nomme la métabolomique. Elle permet d'avoir un cliché instantané des métabolites présents dans la cellule et de les quantifier. Il existe pour cela deux techniques : la spectrométrie de masse et la résonance magnétique nucléaire (NMR), cette dernière étant plus sensible à des petites quantités de métabolites tandis que les études menées par NMR sont hautement reproductibles (Emwas 2015 ; Markley et al. 2017). Lors de ces analyses, les résultats sont donnés sous forme de spectres, qui doivent ensuite être analysés via des méthodes bioinformatiques, en faisant appel à des banques de données de spectres de composés purs. Ces méthodes bioinformatiques sont un domaine en pleine expansion (Johnson et al. 2015 ; Meier et al. 2017). Il est également possible de réaliser des

analyses métabolomiques sur des communautés d'organismes. On parle alors de meta-metabolomique (Turnbaugh et al. 2007).

A.3.2.2 Les réactions biochimiques et les enzymes

Les métabolites réagissent entre eux lors de réactions biochimiques. Ces réactions biochimiques ont comme point de départ un ou plusieurs composés chimiques nommés substrats. Les métabolites en sortie de la réaction sont nommés produits. Les réactions biochimiques ayant lieu dans un milieu vivant, les concentrations de substrats, le pH et la température (entre autres) varient, ce qui influe sur la vitesse ou le sens des réactions biochimiques, chacune possédant des caractéristiques qui lui sont propres. Un catalyseur est un composé chimique permettant d'accélérer, voire d'inverser, une réaction chimique. Bien que prenant part à la réaction, celui-ci n'est pas consommé lors de la réaction. Dans le cadre du métabolisme, les catalyseurs principaux sont les enzymes.

Les enzymes sont des protéines dotées de propriétés catalytiques. On parle alors de biocatalyseurs. Celles-ci permettent d'accélérer jusqu'à 17 fois (Neet 1998) la vitesse des réactions biochimiques. Il existe également des ribozymes qui ne sont pas des protéines mais des molécules d'ARN (Müller et al. 2016). C'est le cas des ribososmes qui catalysent la traduction des ARNm en protéines.

Les enzymes sont classées selon une systématique appelée nomenclature EC (de l'anglais : *Enzyme Commission number* (EC number)). La nomenclature EC a été proposée en 1958 par Dixon et Webb et adoptée par l'Union internationale de biochimie et de biologie moléculaire (IUBMB, de l'anglais : *International Union for Biochemistry and Molecular Biology*) en collaboration avec l'IUPAC (McDonald, Boyce et Tipton 2015). Chaque nouvelle enzyme doit être présentée au comité de nomenclature de l'IUBMB afin de lui attribuer une nomenclature EC. La nomenclature EC est unique à chaque enzyme et est composée de 4 nombres séparés par un point (exemple : Figure A.XII). Les critères de classification sont expliqués dans le tableau A.10.

1er nombre	2enombre	2e nombre	4e nombre	Exemple
1 (Oxidoreductases)	Groupe chimique réducteur	Groupe oxydant	Numéro de série	Oxydase
2 (Transferases)	Groupe transféré	Groupe transféré		Kinase
3 (Hydrolases)	Nature de la liaison hydrolysée	Nature du substrat		Lipase, Peptidase
4 (Lyases)	Nature de la liaison cassée	Groupe éliminé		Décarboxylase
5 (Isomerases)	Type d'isomérisme	Type de substrat		Isomérase, mutase
6 (Ligases)	Nature du lien formé	Type de substrat		Ligase, synthase
7 (Translocase)	Molécule translocatée	Fournisseur d'énergie		Transporteur

Tableau A.10 – Critères de classification des enzymes par la nomenclature EC.

D'après McDonald, Boyce et Tipton (2015) et de `https://www.qmul.ac.uk/sbcs/iubmb/enzyme/`

$$\boxed{EC\ 3.2.1.1}$$

Figure A.XII – Exemple de nomenclature EC pour l'enzyme amylase. Le 3 signifie qu'elle appartient aux hydrolases. Le 2, aux enzymes de type glycosidase. Le dernier 1 est le numéro de série des enzymes de ce type, l'amylase étant la première.

En ce qui concerne les noms d'enzymes, ils doivent être le moins ambigu et le plus précis possible. Pour cela, il existe des règles précises consultables sur le site ExploEnz (`http://www.enzyme-database.org/rules.php`). Ce site présente également la liste d'enzymes validée par l'IUBMB (McDonald et al. 2007 ; McDonald, Boyce et Tipton 2009).

Nous avons décrit tous les acteurs du métabolisme et leurs rôles. Mais afin de mieux comprendre le métabolisme, il est maintenant question de rassembler ces acteurs dans un même concept, les voies métaboliques.

A.3.2.3 Les voies métaboliques

Les voies métaboliques sont un concept réunissant les différents acteurs du métabolisme précédemment évoqués. Il s'agit d'une suite de réactions biochimiques impliquant des métabolites qui sont catalysés par des enzymes ou rybozymes, ARNs capables de catalyser une réaction chimique. Ces voies métaboliques sont représentées comme un chemin de réaction. Elles sont présentes sous deux formes (Figure A.XIII) : (i) sous forme linéaire, où un ou plusieurs métabolites substrats sont le point de départ de la voie métabolique et où un ou plusieurs métabolites sont produits à la fin. Les métabolites produits lors des différentes réactions de la voie mais qui ne sont pas les métabolites produits finaux de la voie métabolique sont appelés métabolites intermédiaires. Une voie linéaire peut se séparer en deux ou réunir une série de réactions. On parle alors de voie métabolique ramifiée car les branches de la voie ont lieu en parallèle. (ii) Sous forme cyclique, où tous les métabolites sont intermédiaires. Les molécules, substrats et produits rentrent ou sortent du cycle lors de réactions biochimiques.

Une même voie métabolique, bien que produisant les mêmes produits *in fine*, peut ne pas présenter le même cheminement de réactions. On parle alors de variante d'une voie métabolique (Romero et al. 2005). Cinq variantes de la voie de la glycolyse ont par exemple été mises en évidence, chacune ayant différents métabolites servant de substrat et ayant lieu dans différents taxons. Les voies métaboliques peuvent appartenir au métabolisme primaire ou secondaire mais peuvent également être amphiboliques (contenant des réactions biochimiques à la fois anaboliques et cataboliques). Certaines voies métaboliques sont également qualifiées de centrales car elles fournissent l'énergie de toute la cellule, ainsi que d'autres éléments biologiques indispensables. On citera comme voie métabolique centrale notamment : la glycolyse, la néoglucogenèse, la voie d'Entner-Doudoroff et la voie des pentoses phosphates (Koonin et Galperin 2003). Les voies métaboliques ne sont pas indépendantes. Les métabolites intermédiaires utilisés dans une voie peuvent être utilisés dans une autre voie métabolique. L'ensemble des voies métaboliques

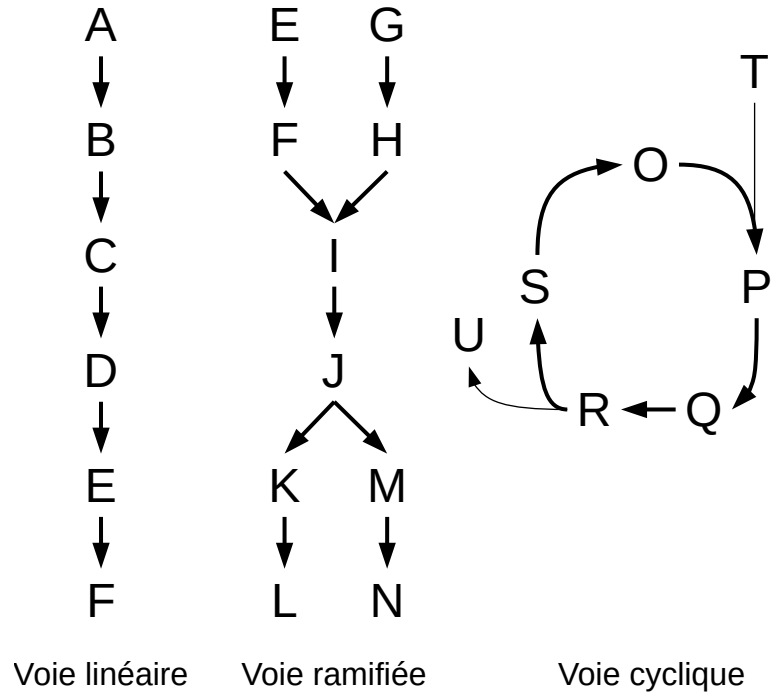


FIGURE A.XIII – Les différentes formes de voies métaboliques. Les lettres représentent les métabolites, les flèches représentent les réactions. Les métabolites A, E et F sont des métabolites substrats et les métabolites F, L, N sont des métabolites produits. Dans la voie cyclique, le métabolite T est introduit dans le cycle afin que celui-ci continue à produire le métabolite U.

D'après WILLEY et al. (2017)

ainsi que les réactions métaboliques isolées de toutes voies forment le réseau métabolique de la cellule (Figure A.XIV).

A.3.3 Les bases de données dédiées aux acteurs du métabolisme

Il existe de nombreuses bases de données dédiées aux acteurs du métabolisme. Certaines proposent des informations sur l'ensemble des composants du métabolisme (métabolites, enzymes, gènes codant ces enzymes, réactions et voies métaboliques) pour l'ensemble des domaines du vivant, tandis que d'autres sont spécialisées pour un seul composant ou une seule espèce. Dans cette partie, nous présentons plus particulièrement neufs d'entre elles.

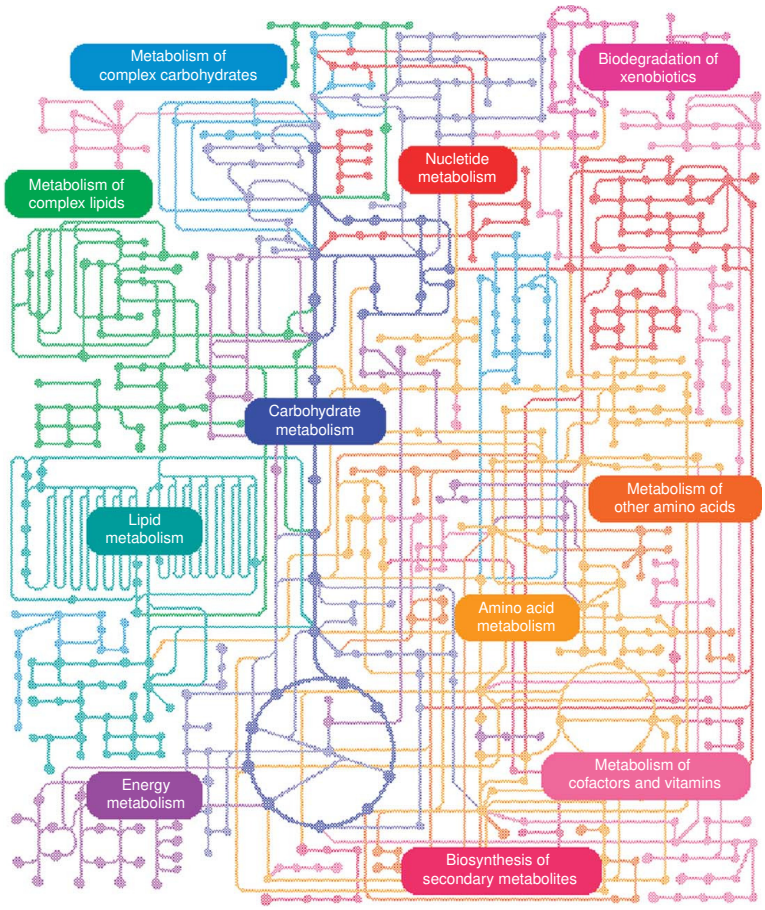


FIGURE A.XIV – Schéma du réseau métabolique de *Escherichia coli*.

Source: LENGAUER et HARTMANN (2007)

A.3.3.1 MetaCyc & BioCyc

MetaCyc (`https://metacyc.org/`) est une base de données de voies métaboliques mises en évidence expérimentalement, et provenant de tous les domaines du vivant (CASPI et al. 2018 ; CASPI et al. 2014 ; KARP et al. 2000). Chaque voie métabolique est décrite (Figure A.XV, page 63) et est associée à la littérature scientifique la décrivant et à un ou plusieurs taxons où celle-ci est présente. Les métabolites, réactions, enzymes et gènes associés à ces voies sont aussi inclus dans MetaCyc. Au moment de l'écriture de ce manuscrit, celle-ci est en version 23.0 (avril 2019, `https://metacyc.org/release-notes.shtml`) et contient 2 722 voies métaboliques présentes dans 3 009 organismes. Ces voies sont composées de 15 767 réactions catalysées par 12 267 enzymes différentes et associées avec 15 655 composés chimiques.

BioCyc (`https://biocyc.org/`) est une collection de bases de données génomes/voies métaboliques nommées PGDBs, de l'anglais *Pathway/Genome DataBases* (Karp et al. 2017 ; Caspi et al. 2016). Une PGDB associe le génome d'un organisme aux voies métaboliques, aux réactions, aux enzymes, aux gènes et aux métabolites identifiés *in silico* dans le génome et/ou expérimentalement dans son métabolisme.

BioCyc intègre MetaCyc dans sa totalité, cette dernière étant elle-même structurée sous un format PGDB adéquat. La version 23.0 (avril 2019, `https://biocyc.org/release-notes.shtml`) de BioCyc contient 14 286 PGDBs de procaryotes. Les PGDBs de BioCyc sont divisées en trois niveaux : (i) le premier niveau (*Tier 1*) distingue les PGDBs ayant bénéficié au minimum d'un an d'annotation manuelle basée sur la littérature. Ce sont les PGDBs les plus précises dans leur information. La seule PGDB procaryote de niveau *Tier 1* est celle d'*Escherichia coli* K-12 substr. MG1655 (EcoCyc, Keseler et al. 2013). MetaCyc est également considéré comme une PGDB de premier niveau, car annotée manuellement. (ii) Les PGDBs de deuxième niveau (*Tier 2*) ont, quant à elles, été générées en utilisant le logiciel Pathway Tools et plus précisément sa composante nommée PathoLogic sur des génomes provenant de GenBank ou de RefSeq. Les faux positifs (c.-à-d. voies détectées mais absentes du génome) sont ensuite corrigés manuellement. Cet effort de curation représente un temps de travail d'un expert d'une durée moyenne de un à quatre mois par PGDB. Ce niveau 2 est actuellement composé de 40 organismes. (iii) Les PGDBs de troisième niveau (*Tier 3*) sont générées automatiquement grâce à PathoLogic à partir de génomes issus de GenBank pour la plupart. Aucune annotation manuelle n'est apportée à ces PGDBs. Dans sa version 23.0, BioCyc contient 14 245 organismes procaryotes de troisième niveau.

La licence de consultation et d'utilisation de ces deux bases est la suivante : la consultation de MetaCyc en ligne est libre et ouverte à tous. Il en va de même pour les PGDBs de BioCyc datées de plus de deux ans. Les universitaires peuvent accéder librement à toutes les fonctionnalités et données des différentes bases de données (deux ans d'ancienneté pour les PGDBs BioCyc) après demande d'une licence. Les non-universitaires doivent s'acquitter d'une licence annuelle afin d'avoir accès à Pathway Tools, à MetaCyc et aux PGDBs de BioCyc .

Le logiciel Pathway Tools (Karp et al. 2015 ; Karp, Latendresse et Caspi 2011) est développé en parallèle de MetaCyc et BioCyc. Il permet de construire, mettre à jour, visualiser et analyser les PGDBs. Pour générer une PGDB, Pathway Tools se base sur l'annotation fonctionnelle des gènes. L'annotation de ces gènes ou de leurs produits permet la reconstruction des réactions métaboliques présentes dans le génome et *in fine* la reconstruction

des voies métaboliques de l'organisme. PathoLogic est le composant de Pathway Tools en charge la génération des PGDBs. Il est également possible d'accéder aux données par une API (interface de programmation d'application de l'anglais : *Application Programming Interface*) disponible dans différents langages informatiques (Krummenacker et al. 2005). L'interface web de Pathway Tools contient des outils permettant l'exploitation des données génomiques et métaboliques de BioCyc (Paley, Latendresse et Karp 2012 ; Travers et al. 2013).

La base de donnée MetaCyc via l'utilisation de Pathway Tools a permis de générer des PGDBs qui ont été utilisées lors des travaux présentés dans cette thèse.

A.3.3.2 MicroCyc

MicroCyc (`http://www.genoscope.cns.fr/agc/microscope/metabolism/microcyc.php`) est une collection de PGDBs crée dans le cadre du projet Microscope (Vallenet et al. 2009 ; Vallenet et al. 2017). Ces PGDBs sont elles aussi générées par Pathway Tools mais bénéficient également (i) d'un processus de réannotation de la part du LABGEM, (ii) d'un processus de détection d'enzyme via le logiciel PRIAM (Claudel-Renard et al. 2003) et (iii) d'un processus de curation par des biologistes via la plate-forme MaGe (Vallenet et al. 2006).

De par ses qualités, MicroCyc a également été intégré lors des travaux présentés dans cette thèse.

A.3.3.3 KEGG

KEGG (`https://www.genome.jp/kegg/`, the Kyoto Encyclopedia of Genes and Genomes)) est une base de données contenant des données métaboliques (Kanehisa et al. 2019 ; Kanehisa et al. 2017 ; Kanehisa et Goto 2000). KEGG PATHWAY contient les voies métaboliques représentées sous forme de carte métabolique (Figure A.XVI) pouvant varier selon l'organisme d'intérêt.

Ces voies sont organisées selon des catégories (KEGG BRYTE) et incluent des composés chimiques (KEGG COMPOUND) et des réactions (KEGG REACTION) catalysées par des enzymes (KEGG ENZYME). Toutes ces informations sont déduites des informations d'annotation de gènes (KEGG GENES) des génomes présents dans KEGG (KEGG ORGANISM). Les gènes orthologues (gènes homologues présents dans deux espèces différentes issus d'un même gène d'un ancêtre commun direct ayant subi une spéciation) sont reliés via les KO (KEGG ORTHOLOGS). KEGG possède également de

nombreux outils permettant l'exploration des données issues du séquençage à haut débit (Kanehisa et al. 2017 ; Kanehisa et al. 2019). KEGG possède un modèle économique avec abonnement : les données sont accessibles librement via l'interface web ou via l'API, mais les fichiers de données ne sont accessibles qu'en souscrivant à cet abonnement.

A.3.3.4 Comparaison entre MetaCyc/BioCyc et KEGG

Les voies métaboliques de MetaCyc/BioCyc et de KEGG présentent quelques différences. Les voies métaboliques de KEGG sont présentées sous forme de carte indiquant les différentes réactions possibles dans la voie métabolique et il est possible d'afficher seulement les réactions possibles dans un organisme précis. Les voies MetaCyc sont quant à elles sous forme de schémas et sont associées à un taxon. Une voie peut donc être associée à une espèce, un genre, voire à un domaine. Le processus de curation est aussi différent : l'annotation des voies métaboliques de MetaCyc est réalisée par l'importation de données extérieures et de données de la littérature scientifique. Les données sont synthétisées et la qualité vérifiée, validant ainsi l'intégration de la voie dans MetaCyc. Dans KEGG, les voies métaboliques sont annotées par des experts, mais aucune littérature n'est rattachée aux différentes voies. Altman et al. (2013) et Karp et al. (2019) ont réalisé une comparaison entre MetaCyc et KEGG concernant la partie traitant des voies métaboliques. Le tableau A.11 présente une comparaison des informations contenues dans BioCyc et KEGG.

	MetaCyc/BioCyc	**KEGG**
Nombre de voies métaboliques	2 722	534
Nombre de réactions métaboliques	15 767	11 040
Nombre de composés chimiques	15 655	18 537
Nombre d'enzymes	12 267	7 564
Nombre de gènes	12 622	29 545 122
Nombre de génomes	14 728	5 963

Tableau A.11 – Comparaison du nombre de voies métaboliques, de réactions métaboliques, de composés chimiques, d'enzymes, de gènes et de génomes présents dans les bases de données MetaCyc/BioCyc et KEGG. Basé sur MetaCyc/BioCyc version 23.0 (29/04/2019) et sur KEGG release 90.1 (01/05/2019).

A.3.3.5 PATRIC

PATRIC (`https://www.patricbrc.org/`, Pathosystems Resource Integration Center) a été conçu dans le but de soutenir les travaux de recherche sur les maladies infectieuses bactériennes, en fournissant des informations sur les bactéries infectieuses et les outils permettant de les analyser (Wattam et al. 2017). PATRIC contient des données génomiques annotées et associées à des métadonnées écologiques et techniques, de transcriptomiques, d'interactions protéine-protéine et de structures 3D des protéines. Le tout est accompagné d'outils permettant d'analyser ces informations. PATRIC permet d'obtenir une information métabolique grâce à son annotation des génomes permettant d'obtenir les voies métaboliques présentes dans chacun d'entre eux. Actuellement, PATRIC contient également des données pour des bactéries non pathogènes et est la base de données rassemblant le plus de génomes d'organismes (au 20/09/2019 : 252 471 génomes bactérien et 3 717 d'Archaea)

A.3.3.6 WikiPathways

WikiPathways (`https://www.wikipathways.org/`) est une base de données de voies métaboliques participative (Slenter et al. 2018). Celle-ci suit le modèle de Wikipédia, ce qui lui permet d'être ouvert à la modification par tous. Chacun des nouveaux apports est vérifié sur la forme afin que chaque voie métabolique soit associé avec de la littérature confirmant sa présence et sa composition. Il est également possible pour les utilisateurs de vérifier les données présentes. WikiPathways est dédiée aux voies métaboliques de 25 organismes (version 20190610) dont trois bactéries : *Bacillus subtilis*, *Escherichia coli*, *Mycobacterium tuberculosis*. Les voies métaboliques sont complétées par les gènes, les réactions, les enzymes et les métabolites les composant. Les données peuvent être consultées librement en ligne ou via une API.

A.3.3.7 FAPROTAX

Functional Annotation of Prokaryotic Taxa (FAPROTAX, `https://pages.uoregon.edu/slouca/LoucaLab/archive/FAPROTAX/lib/php/index.php` est une base de données associant fonctions métaboliques et écologiquement pertinentes à des taxons (Louca, Parfrey et Doebeli 2016). Elle contient 82 fonctions pour plus de 7 600 taxons. Ces annotations proviennent de la littérature traitant des procaryotes tel que le journal IJSEM, *The Prokaryotes* (Rosenberg et al. 2013) ou le *Begrey's Manual of Systematic Bacteriology*.

A.3.3.8 IJSEM phenotypic database

International Journal of Systematic and Evolutionary Microbiology phenotypic database (`https://figshare.com/articles/International_Journal_of_Systematic_and_Evolutionary_Microbiology_IJSEM_phenotypic_database/4272392`) est une base de données contenant des données phénotypiques, métaboliques et de tolérance environnementale des procaryotes (Barberán et al. 2017). Elle contient 16 fonctions rassemblées manuellement à partir du journal IJSEM.

FAPROTAX et IJSEM phenotypic database ont été utilisés lors des travaux de cette thèse.

A.3.3.9 BRENDA

BRENDA (`https://www.brenda-enzymes.org/`, BRaunschweig ENzyme DAtabase) est une base de données d'enzymes et de métabolites de très haute qualité (Jeske et al. 2019 ; Schomburg et al. 2017). Elle contient également 169 voies métaboliques (version 2019.01). Les informations présentes dans BRENDA sont extraites de la littérature scientifique, ce qui permet d'obtenir une information de grande qualité. Il est à noter que les noms d'organismes associés ne sont pas mis à jour si un organisme ou une souche subissent un reclassement taxonomique.

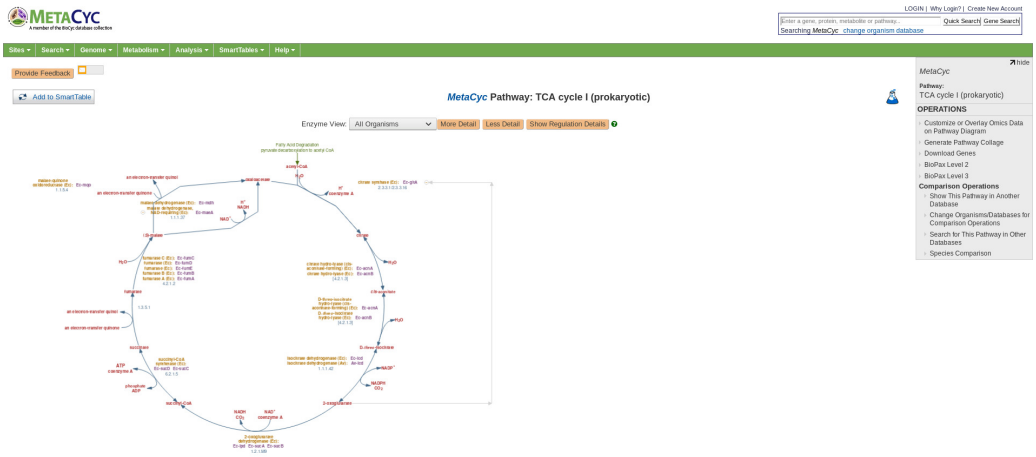
A.3.3.10 UniProt

UniProt (UNIversal PROTein reference, `https://www.uniprot.org/`) est une collaboration entre l'EMBL-EBI (European Molecular Biology Laboratory – European Bioinformatics Institute), le SIB (Swiss Institute of Bioinformatics) et le PIR (Protein Information Ressource) et fournit aux scientifiques une source librement accessible et de haute qualité de séquences de protéines (Consortium 2019). UniProt est constituée de plusieurs parties interconnectées : (i) UniProtKB (Uniprot Knowledgebase), elle-même divisée en deux parties : Swiss-Prot, base de données de protéines annotées manuellement depuis la littérature, et TrEMBL, bases de données de protéines annotées automatiquement à partir de Swiss-Prot, (ii) UniRef (UniProt Reference Clusters) qui sont des ensembles de séquences de protéines similaires. Trois résolutions sont disponibles : 100%, 90% et 50%. (iii) UniParc (UniProt Archive) est une archive de toutes les protéines pouvant être extraites des banques de données publiques, (iv) Proteomes, qui réunit de multiples ensembles de protéines exprimées par un organisme précis. Dans le cas d'une étude métabolique, UniProt permet d'identifier les enzymes, protéines possédant un pouvoir catalytique, et de les lier à leur réaction biologique. Bien

que les enzymes présentes dans UniProtKB puissent être associées à des voies métaboliques, celles-ci proviennent de UniPathway (Morgat et al. 2012), ressource qui réunissait et annotait manuellement des voies métaboliques, mais celle-ci ne semble plus active.

A.3.3.11 Pérennité et modèles économiques des bases de données métaboliques

Le tableau 1 (page 104) de l'article MACADAM reprend les principales caractéristiques des bases de données BioCyc, KEGG et PATRIC. Chacune d'entre elles présente des qualités particulières. BioCyc permet aux académiques d'utiliser les voies métaboliques et le logiciel Pathway Tools sans restrictions et possède un grand nombre de voies métaboliques. KEGG présente l'intégralité des voies métaboliques connues d'un organisme sur une carte permettant une compréhension immédiate du réseau métabolique d'un organisme. La base de données PATRIC contient un nombre d'organismes procaryotes très important. Mais la question de la pérennité de ces bases de données et de la fiabilité des informations contenues dans celles-ci se pose. En effet, la curation de ces bases de données demande énormément de temps et de ressources pour que l'information biologique qu'elles contiennent continue à être actualisée. KEGG et Biocyc ont basculé vers un modèle à abonnement en 2011 (`https://www.kegg.jp/kegg/docs/plea.html`) et en 2016 (`https://biocyc.org/news001-subscriptions.shtml`, Karp et al. 2017 ; Caspi et al. 2018) respectivement. PATRIC et MicroCyc restent libres d'accès. La pérennité des bases de données et notamment de la durée de leur financement est une question centrale des banques de données biologiques actuelles (Parkhill, Birney et Kersey 2010 ; Bourne, Lorsch et Green 2015 ; Reiser et al. 2016). Ces considérations peuvent expliquer les limitations imposées par les banques de données de ne pas rendre leurs informations biologiques téléchargeables librement.



This view shows enzymes only for those organisms listed below, in the list of taxa known to possess the pathway. If an enzyme name is shown in bold, there is experimental evidence for this enzymatic activity.

Some taxa known to possess this pathway include : Azotobacter vinelandii, Escherichia coli K-12 substr. MG1655

Expected Taxonomic Range: Archaea, Bacteria

Synonyms: tricarboxylic acid cycle, citric acid cycle, Krebs cycle, Szent-Gyorgyi-Krebs cycle

Superclasses: Generation of Precursor Metabolite and Energy → TCA cycle

Summary:

General Background

The TCA pathway is a catabolic pathway of aerobic respiration that generates both energy and reducing power. In addition, it is also the first step in generating precursors for biosynthesis. The pathway is very common, and a variation of it exists in practically all aerobic living organisms [Krebs37, Krebs38, Krebs45].

The input to the cycle is acetyl-CoA, an activated form of acetate that is generated by the degradation of carbohydrates, fats and proteins. A common source of acetyl-coA is pyruvate, which is generated by glycolysis and converted to acetyl-CoA by the pyruvate dehydrogenase complex.

In every turn the TCA cycle converts one molecule of acetyl-CoA into two CO_2 molecules, reduces a total of four molecules of either NAD^+, $NADP^+$, or quinone to NADH, NADPH and quinol, respectively, and phosphorylates one molecule of GDP to GTP.

The reduced molecules of NADH/NADPH/quinol that are formed by the TCA cycle serve as electron donors for oxidative phosphorylation (see for example aerobic respiration I (cytochrome c)). In that process the electrons flow to a terminal acceptor, powering on their way proton pumps that trasport protons across the cytoplasmic or mitochondrial membranes, generating a proton motive force (PMF). As the protons return to their original location, they power ATPase enzymes that phosphorylate ADP molecules to ATP. The total energy gained from the complete breakdown of one molecule of glucose by glycolysis, the TCA cycle, and oxidative phosphorylation equals about 30 ATP molecules in eukaryotes.

The name of the TCA (short for tricarboxylic acid) cycle is derived from the fact that the first step in the pathway is attachment of acetyl-coA to citrate, an acid with three carboxylate groups. The pathway is also known as the citric acid cycle, and as the Szent-Gyorgyi-Krebs cycle (or just the Krebs cycle), named after the scientists who described it.

Some organisms possess a truncated version of the TCA cycle, known as the glyoxylate cycle, that converts acetyl-CoA to biosynthetic intermediates without the loss of CO_2 [Walsh84, Nimmo87, Holms87].

About This Pathway

This is a common variation of the pathway that occurs in many bacteria and archaea. There are a few small differences between this prokaryotic version of the cycle and the version found in most eukaryotes (see TCA cycle II (plants and fungi)). In this pathway, an NADP+-dependent enzyme (EC 1.1.1.42) catalyzes the dehydrogenation of D-threo-isocitrate to 2-oxoglutarate, while eukaryotes employ an NAD+-dependent enzyme (EC 1.1.1.41). Another difference is that while in most eukaryotes the conversion of (S)-malate to oxaloacetate is catalyzed only by an NAD-dependent enzyme (EC 1.1.1.37), prokaryotes that employ this variation of the TCA cycle possess an alternative quinone-dependent enzyme (EC 1.1.5.4).

While the pathway is most common in heterotrophic bacteria and arachaea, there is evidence for its presence in some facultatively autotrophic archaea when growing under heterotrophic conditions.

In *Escherichia coli*, when acetate is the carbon source, citrate synthase is rate-limiting for the TCA cycle [Walsh85a, Walsh87].

Superpathways: superpathway of glyoxylate bypass and TCA, superpathway of glycolysis, pyruvate dehydrogenase, TCA, and glyoxylate bypass

Variants: partial TCA cycle (obligate autotrophs), TCA cycle II (plants and fungi), TCA cycle III (animals), TCA cycle IV (2-oxoglutarate decarboxylase), TCA cycle V (2-oxoglutarate:ferredoxin oxidoreductase), TCA cycle VII (acetate-producers), TCA cycle VIII (Helicobacter)

Unification Links: AraCyc:TCA, EcoCyc:TCA

Credits:
Revised 19-Dec-2011 by Caspi R, SRI International

References

Baldwin81a: Baldwin JE, Krebs H (1981). **"The evolution of metabolic cycles."** Nature 291(5814):381-2. PMID: 7242661

Holms87: Holms WH (1987). **"Control of flux through the citric acid cycle and the glyoxylate bypass in Escherichia coli."** Biochem Soc Symp 1987;54;17-31. PMID: 3332993

Krebs37: Krebs HA, Johnson WA (1937). **"Acetopyruvic acid (try-diketovaleric acid) as an intermediate metabolite in animal tissues."** Biochem J 31(5);772-9. PMID: 16746397

Krebs38: Krebs HA, Salvin E, Johnson WA (1938). **"The formation of citric and α-ketoglutaric acids in the mammalian body."** Biochem J 32(1);113-7. PMID: 16746585

Krebs45: Krebs HA, Eggleston LV (1945). **"Metabolism of acetoacetate in animal tissues. 1."** Biochem J 39(5);408-19. PMID: 16747930

Nimmo87: Nimmo HG, Borthwick AC, el-Mansi EM, Holms WH, MacKintosh C, Nimmo GA (1987). **"Regulation of the enzymes at the branchpoint between the citric acid cycle and the glyoxylate bypass in Escherichia coli."** Biochem Soc Symp 1987;54;93-101. PMID: 3333001

Walsh84: Walsh K, Koshland DE (1984). **"Determination of flux through the branch point of two metabolic cycles. The tricarboxylic acid cycle and the glyoxylate shunt."** J Biol Chem 1984;259(15);9646-54. PMID: 6378912

Walsh85a: Walsh K, Koshland DE (1985). **"Characterization of rate-controlling steps in vivo by use of an adjustable expression vector."** Proc Natl Acad Sci U S A 1985;82(11);3577-81. PMID: 3889909

Walsh87: Walsh K, Schena M, Flint AJ, Koshland DE (1987). **"Compensatory regulation in metabolic pathways--responses to increases and decreases in citrate synthase levels."** Biochem Soc Symp 1987;54;183-95. PMID: 3332995

FIGURE A.XV – Fiche du cycle de Krebs (*TCA cycle* en anglais) procaryote dans MetaCyc.

Source: MetaCyc

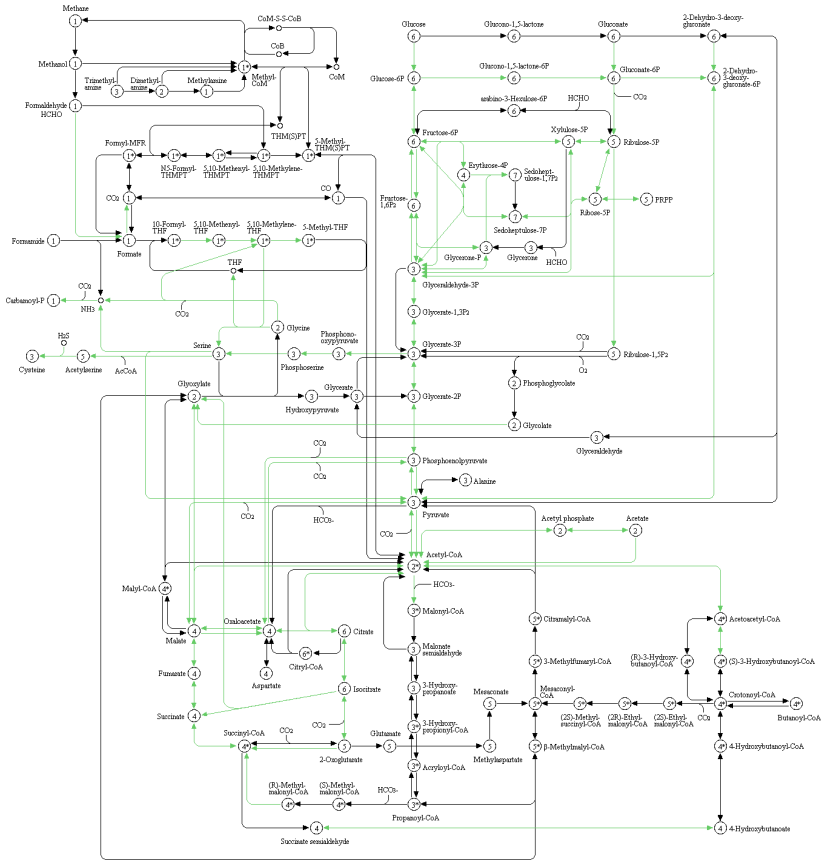


FIGURE A.XVI – Ensemble des voies du métabolisme du carbone présentes chez *Escherichia coli* K-12 MG1655. Les flèches vertes indiquent que le gène codant l'enzyme catalysant la réaction a été identifié et que sa séquence ainsi que la séquence d'acides aminés de l'enzyme est disponible.

Source: KEGG

A.3.4 Inférence fonctionnelle d'une communauté procaryotique

Dans cette partie nous détaillerons plus particulièrement les avantages et les limites d'outils dédiés à l'analyse fonctionnelle des communautés complexes à partir de données de séquençage de gènes marqueurs (typiquement le gène de l'ARNr 16S) ou à partir de données de séquençage métagénomique.

L'inférence est définie par le centre national de ressources textuelles et lexicales comme étant « une opération qui consiste à admettre une proposition en raison de son lien avec une proposition préalable tenue pour vraie ». Nos travaux portent sur l'inférence fonctionnelle, que l'on peut définir comme étant une méthode qui consiste à admettre une proposition de potentiel fonctionnel d'un organisme ou d'une communauté en raison de la seule présence de séquences connues portant cette fonction dans le génome de cet organisme ou des organismes de cette communauté. Le potentiel fonctionnel d'un organisme ou d'une communauté est défini comme ses capacités métaboliques, c'est-à-dire la présence du ou des gènes permettant la réalisation de la fonction. Cela implique qu'en l'absence de données d'expression de ces gènes l'organisme peut ne pas réaliser la fonction malgré la présence des gènes en question.

Des outils d'inférence fonctionnelle ont été développés pour déterminer le potentiel fonctionnel à partir des données issues de séquençage métagénomique (tel que HUMAnN2) ou pour prédire le potentiel fonctionnel à partir de données de séquençage de gènes marqueurs avec une étape d'identification ou d'affiliation taxonomique (PICRUSt, Tax4Fun et PAPRICA). Ces outils présentent des avantages et des limites inhérents à la méthode de séquençage choisie ou provenant de la source des données taxonomiques et fonctionnelles sur lesquelles les outils reposent.

Les outils réalisant une inférence fonctionnelle à partir de séquençage de gène marqueurs utilisent une base de données de profils fonctionnels de référence pré-calculés et associés à une taxonomie ou à une séquence de gène marqueurs. Ces outils nécessitent comme données d'entrée une table d'abondance associant une abondance avec une séquence ou une affiliation taxonomique. Cette technique présente certains avantages (Douglas, Beiko et Langille 2018) : (i) la puissance de calcul nécessaire pour l'analyse fonctionnelle est moindre, (ii) la séquence du gène de l'ARNr 16S doit être incluse lors de la description d'une nouvelle espèce (Stackebrandt et al. 2002). Les bases de données génomiques contiennent donc des souches dont seule la séquence du gène de l'ARNr 16S est disponible, tandis que le génome complet n'est pas disponible. Le gène de l'ARNr 16S permet d'accéder à une plus grande diversité taxonomique. Mais l'inférence fonctionnelle à partir de

gènes marqueurs repose sur la prédiction du potentiel fonctionnel : le potentiel n'est pas observé directement, mais repose sur des profils pré-calculés qui peuvent ne pas être représentatifs de l'organisme présent dans le milieu. La quantification des fonctions présentes dans le potentiel fonctionnel par cette méthode est également ardue car, pour le gène de l'ARNr 16S, son nombre de copies varie entre les différentes espèces et les souches d'une même espèce (Acinas et al. 2004 ; Větrovský et Baldrian 2013). La génération des profils fonctionnels est également limitée aux organismes dont le génome et/ou les fonctions sont disponibles dans les banques de données.

L'inférence fonctionnelle à partir de lectures provenant d'un séquençage métagénomique ne nécessite pas d'affiliation taxonomique préalable. Les lectures sont soit directement assignées à une fonction, soit assemblées puis assignées à une fonction. Cette méthode permet d'observer le potentiel fonctionnel tel qu'il est présent dans le milieu et non de le prédire à partir des gènes marqueurs par exemple. L'inférence fonctionnelle à partir de lectures issues de séquençage métagénomique est également capable d'obtenir le potentiel fonctionnel d'organismes inconnus ou non cultivés si la séquence de la lecture est présente dans une base de données de protéines. Mais cette approche peut être impossible matériellement et financièrement pour certains milieux. En effet, l'inférence fonctionnelle à partir de lectures métagénomiques demande une profondeur de séquençage entraînant des coûts qui, si l'on associe ceux-ci avec un nombre important d'échantillons, peuvent ne pas être possibles à supporter. Une profondeur de séquençage trop faible lors d'un séquençage métagénomique ne permet que d'entrevoir les fonctions les plus abondantes dans la communauté et ne permet pas d'identifier les fonctions plus rares (Langille et al. 2013). Le séquençage amplicon permet d'obtenir une meilleure profondeur de séquençage, d'englober tous les organismes présents dans le milieu et *in fine* de réaliser une inférence fonctionnelle à moindre coût. Les analyses à mener sur les données de séquençage métagénomiques sont également chronophages et coûteuses (Vincent et al. 2017).

Dans les parties suivantes, nous présenterons plus en détail des outils qui étaient présents au début de notre travail de thèse : (i) PIRCRUSt, (ii) Tax4Fun et (iii) PAPRICA. Ces outils à l'époque présentaient certaines limites que nous décrirons. Nous présenterons également les évolutions de ces outils depuis les débuts de ces travaux de thèse.

A.3.4.1 PICRUSt

PICRUSt (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States, Langille et al. 2013) est un outil d'inférence fonctionnelle permettant de déterminer le potentiel fonctionnel à partir du gène

de l'ARNr 16S. Pour cela, PICRUSt contient un arbre phylogénétique issu de Greengenes (Partie A.1.7.5, page 36) dont les feuilles sont des séquences du gène de l'ARNr 16S. Ces séquences feuilles peuvent être accompagnées de leur génome provenant de IMG/M (Markowitz et al. 2012) et de leurs annotations (KEGG Orthologs, KO) ou n'être composées que d'une séquence de gènes de l'ARNr 16S. Pour ces dernières, un profil fonctionnel est inféré à partir de la reconstruction des états ancestraux des feuilles présentant un génome de référence (Figure A.XVII). PICRUSt contient également, pour les feuilles avec génomes, le nombre de copies du gène de l'ARNr 16S et prédit un nombre de copies pour les feuilles sans génome de référence. L'intégralité de ces données fait partie intégrante de PICRUSt et elles ne doivent pas être générées pas l'utilisateur.

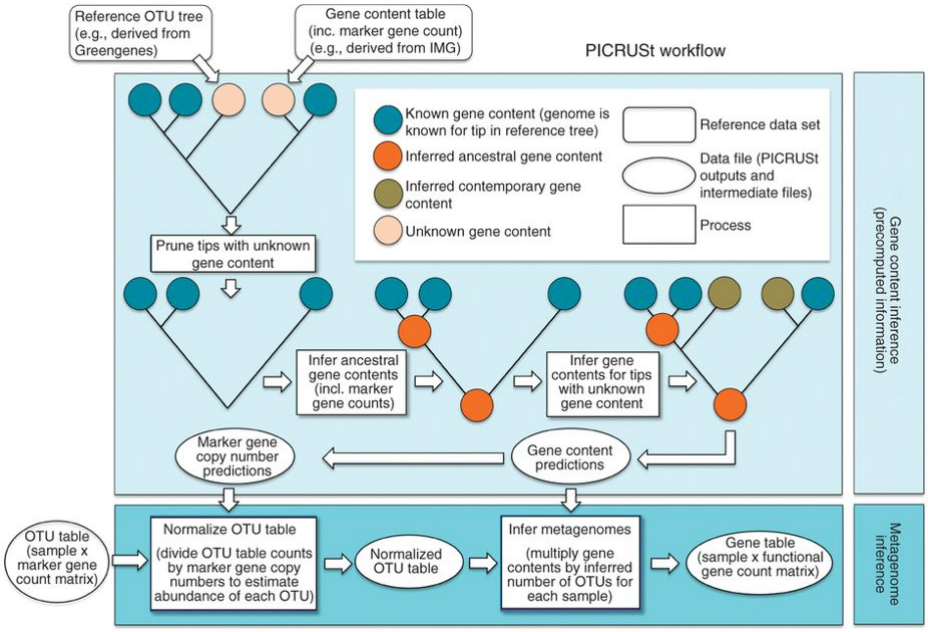


Figure A.XVII – Chaîne de traitement de PICRUSt.

Source: Langille et al. (2013)

Afin d'utiliser PICRUSt, l'utilisateur doit générer une table d'abondance d'OTUs à partir de ses données de séquençage de gènes marqueurs en utilisant QIIME (Caporaso et al. 2010). Pour cela QIIME doit être utilisé pour assigner chacune des OTUs avec un identifiant Greengenes (« Closed-reference OTU picking »). Cet identifiant est utilisé afin de placer l'OTU sur un des profils présents dans PICRUSt.

PICRUSt présente l'avantage de pouvoir inférer un potentiel fonctionnel

pour les OTUs associés à une taxonomie ne présentant pas de génome de référence. Mais PICRUSt présente également plusieurs limites : (i) l'obligation de l'utilisation de Greengenes conjointement avec QIIME pour l'assignation taxonomique des OTUs. La dernière version de Greengenes remontant à mai 2013 (version 13.5, `https://greengenes.secondgenome.com/`), il est probable que des erreurs soient présentes et que celle-ci ne présente pas les derniers organismes mis en évidence. (ii) Les données fonctionnelles de PICRUSt ne sont également plus mis à jour, la base de données KEGG nécessitant une licence depuis 2011 (`https://www.kegg.jp/kegg/docs/plea.html` et `https://github.com/picrust/picrust2/wiki/Frequently-Asked-Questions`). (iii) La prédiction des états ancestraux et des feuilles sans génomes associés dépend de l'arbre phylogénétique Greengenes. Cet arbre contiendrait un taux d'erreur de 17% environ et ne serait pas adapté comme arbre guide (Edgar 2018a).

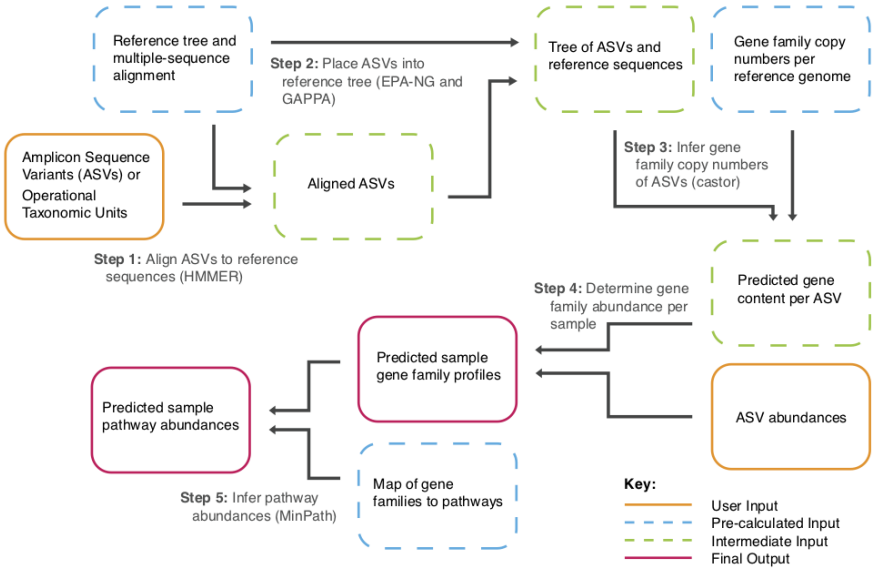


Figure A.XVIII – Chaîne de traitement de PICRUSt2.

Source: Douglas et al. (2019)

Au cours de cette thèse, PICRUSt2 a été développé afin de répondre à certaines de ces limites (pré-version de l'article : Douglas et al. 2019). PICRUSt2 ne nécessite plus la génération d'une table d'OTUs via QIIME en utilisant Greengenes comme référence. Les auteurs recommandent de ne plus utiliser les OTUs mais les ASVs (Partie A.2.1.2, page 44), afin de réaliser la table d'abondance. Chaque ASV de la table d'abondance doit contenir la

séquence représentative du groupe, car contrairement à la première version de PICRUSt qui se reposait sur l'identifiant Greengenes, PICRUSt2 aligne la séquence représentative sur un arbre reconstruit à partir de séquences du gène de l'ARNr 16S. Cet arbre de séquences de gène de l'ARNr 16S est reconstruit à partir de 41 926 séquences de gènes de l'ARNr 16S de génomes provenant d'IMG. Les annotations de familles de gènes associées à ces génomes (sous la forme d'annotation : KEGG Orthologs (version 77.1, juillet 2016), nomenclature EC, COGS, Pfam et TIGRFAM) sont également récupérées et associées à leurs séquences du gène de l'ARNr 16S respectives. Ces annotations sont utilisées afin de former des voies métaboliques avec l'outil MinPath (Ye et Doak 2009) (Figure A.XVIII). Chaque espèce est associée à une taxonomie à l'aide du package R *taxizedb* (`(https://github.com/ropensci/taxizedb`) et du NCBI Taxonomy. Le package R *castor* est utilisé afin d'inférer le nombre de copies de gène en fonction de son placement sur l'arbre (Louca et Doebeli 2018).

A.3.4.2 Tax4Fun

Tax4Fun (Asshauer et al. 2015) est un package R permettant la prédiction de profils fonctionnels à partir de tables d'abondance d'OTU de communauté procaryotique. Tax4Fun repose sur des profils fonctionnels pré-calculés associés à un profil taxonomique issu de SILVA et normalisé par le nombre de copies du gène de l'ARNr 16S présents dans l'organisme (Figure A.XIX). Ces profils fonctionnels sont obtenus à partir des génomes des KEGG organisms (version 64, octobre 2012) et à l'outil UProC (Meinicke 2015). Ils sont composés de KEGG Orthologs (KO). Les profils taxonomiques sont générés en recherchant les séquences du gène de l'ARNr 16S dans les KEGG organisms dans SILVA puis en normalisant l'abondance de ces profils par le nombre de 16S présents dans l'organisme. L'utilisateur doit ensuite procéder à une affiliation taxonomique en utilisant QIIME ou SILVANGS (Quast et al. 2013), tout en utilisant comme référence la base de données de séquences de gène de l'ARNr 16S SILVA. Chaque OTU de la table d'abondance est alors affiliée à une taxonomie SILVA. La table d'abondance est ensuite associée aux profils fonctionnels pré-calculés par l'outil Taxy-Pro (Asshauer et Meinicke 2013 ; Klingenberg et al. 2013). L'utilisateur obtient donc un profil fonctionnel de la communauté procaryotique composé de KO associés avec une abondance. Un avantage de Tax4Fun est que celui-ci est facile d'utilisation car disponible sous forme d'un package R. Cette structure lui permet d'être multiplateforme et, grâce au pré-calcul de ces profils sous forme R, de demander peu de ressources de calcul pour s'exécuter rapidement. Mais Tax4Fun présente également plusieurs limites :

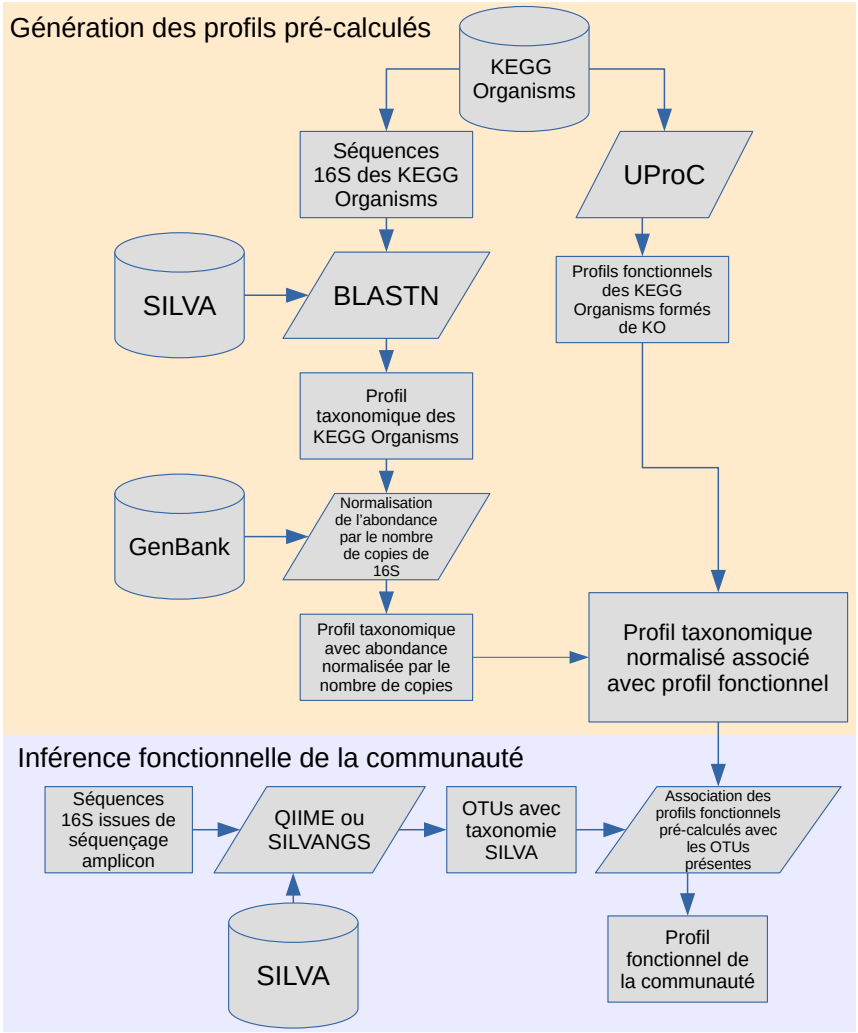


FIGURE A.XIX – Chaîne de traitement de Tax4Fun.

(i) QIIME n'est maintenant plus supporté mais sa seconde version peut être utilisé en modifiant le fichier de sortie ou en utilisant un pipeline alternatif (`https://github.com/peterleary/q2pipeline`). SILVANGS est un service dont l'accès est limité à un certain nombre d'heures par an. L'utilisation de la taxonomie SILVA (version 123, juillet 2015) est obligatoire. (ii) Comme pour PICRUSt, les KO ne sont plus mis à jour depuis le lancement d'une licence pour KEGG (version 64, octobre 2012).

Tax4Fun2 (pré-version de l'article : WEMHEUER et al. 2018) est actuellement en développement. Le principal changement est la génération des profils fonctionnels et des génomes sur lesquels sont basés ces profils. Les génomes proviennent maintenant de RefSeq et présentent une qualité d'as-

semblage minimum de « chromosomes ». Tax4Fun2 utilise toujours UProC mais cette fois-ci avec les KO provenant de la version de juillet 2018 de KEGG (`http://tax4fun.gobics.de/`).

A.3.4.3 PAPRICA

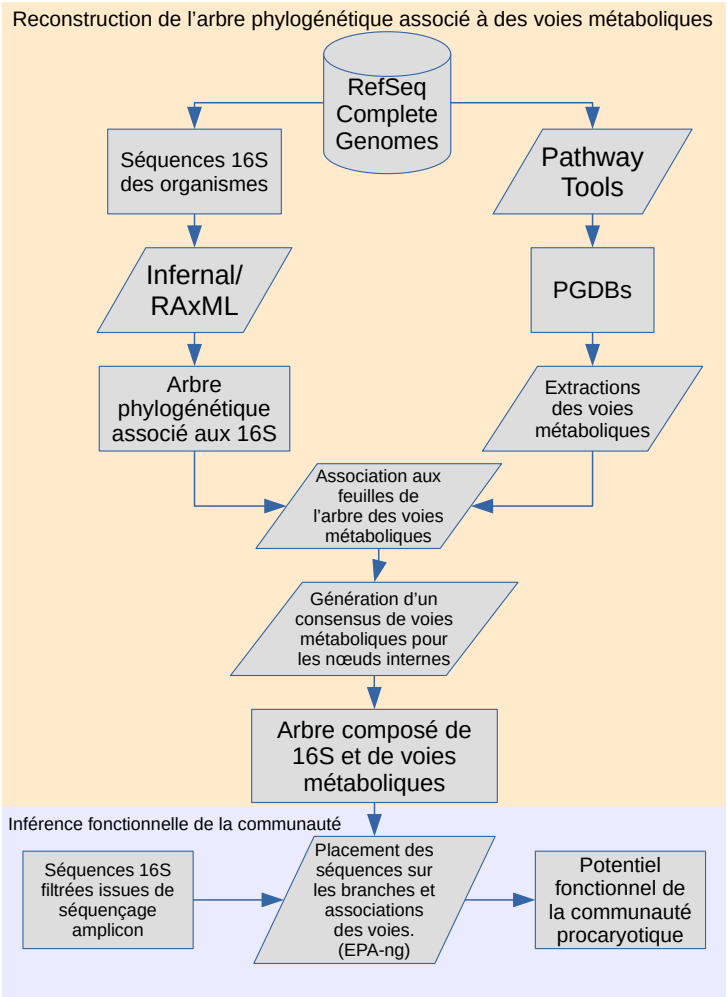


FIGURE A.XX – Chaîne de traitement de PAPRICA.

PAPRICA (PAthway PRediction by phylogenetIC plAcement, BOWMAN et DUCKLOW 2015) est un pipeline permettant de déterminer la structure et le potentiel métabolique d'une communauté procaryotique à partir de séquences de gènes de l'ARNr 16S (Figure A.XX). Pour cela, PAPRICA contient un arbre phylogénétique reconstruit à partir de séquences du gènes de l'ARNr 16S, provenant d'un génome présentant une qualité d'assemblage

« complete genomes ». Chaque feuille de l'arbre est associée à l'un de ces génomes. Une PGDB est ensuite générée via l'utilisation de Pathway Tools à partir de ces génomes provenant de RefSeq et de qualité d'assemblage « complete genome ». Les voies métaboliques composant les PGDBs proviennent de MetaCyc. PAPRICA génère ensuite les voies métaboliques des noeuds intérieurs de l'arbre en se basant sur les voies métaboliques présentes dans les feuilles de l'arbre. Si une même voie est présente 90% du temps dans les noeuds filles, alors la voie est conservée dans les nœuds internes. Afin d'utiliser PAPRICA, l'utilisateur doit fournir les lectures de séquençage, qui sont positionnées sur les branches de l'arbre grâce à l'outil EPA-ng (Barbera et al. 2019). Chaque noeud étant associé avec un certain nombre de voies métaboliques, PAPRICA en déduit ensuite le potentiel fonctionnel de la communauté procaryotique, ainsi que la structure taxonomique du milieu. L'abondance des voies métaboliques est pondérée par le nombre de copies du gène de l'ARNr 16S dans le génome. Les principales limites de PAPRICA sont : (i) la reconstruction d'un arbre phylogénétique à partir de séquences du gène de l'ARNr 16S qui place certains organismes au mauvais endroit, du fait du manque de résolution du gène de l'ARNr 16S dans certains clades. (ii) Le placement des séquences de l'utilisateur peut aussi souffrir de ce manque de résolution.

A.3.4.4 A partir de lectures issues de séquençage métagénomique

L'inférence fonctionnelle à partir de lectures métagénomiques repose sur un principe simple : la nécessité d'associer une séquence avec une annotation fonctionnelle. La séquence peut être une lecture en sortie de séquençage, qui, après contrôle qualité, peut être directement utilisée, comme par exemple dans l'outil nommé HUMAnN2 (Franzosa et al. 2018) que nous utiliserons dans ces travaux de thèse et dont le pipeline est décrit dans la partie B.4, page 128, de l'étude expérimentale. Les lectures peuvent également être assemblées afin de limiter le nombre de séquences à traiter par la suite. Les séquences sont ensuite associées à une fonction en les comparant à des bases de données permettant une affectation fonctionnelle, tels que Pfam, KEGG Orthologs ou encore Uniprot. Pour cela, dans un premier temps, des outils tels que BLAST (Altschul et al. 1990) ou BLAT (Kent 2002) ont été utilisés. Mais ils n'ont pas été conçus dans le cadre d'une utilisation avec un nombre très important de séquences comme produit par un séquençage métagénomique. Sont alors apparus des outils comme USEARCH (Edgar 2010) ou DIAMOND (Buchfink, Xie et Huson 2015) permettant de comparer des séquences beaucoup plus rapidement au détriment de la sensibilité des comparaisons (Bengtsson-Palme 2018). Ces annotations peuvent ensuite

être rassemblées, et il est possible d'inférer des voies métaboliques, notamment via l'outil MinPath (Ye et Doak 2009).

A.3.4.5 Conclusions

Base de données	PICRUSt	Tax4Fun	PAPRICA
Taxonomique	Greengenes (05/2013)	SILVA (07/2015)	NCBI Taxonomy
Fonctionnelle	KEGG (<07/2011)	KEGG (10/2012)	MetaCyc
Génomique	IMG 3.5 (07/2012)	KEGG (10/2012)	RefSeq (complete genome)
Méthode d'inférence	Identifiant Greengenes	Taxonomie SILVA	Placement dans un arbre

Tableau A.12 – Bases de données et méthodes utilisées par les outils d'inférence fonctionnelle disponibles lors du début de ces travaux de thèse.

Le tableau A.12 présente les bases de données et les méthodes utilisées dans les différents outils disponibles au début de cette thèse. PICRUSt et Tax4Fun possédaient une information fonctionnelle et taxonomique obsolètes. Tous les outils d'inférence quelles que soient les données sources (gènes marqueurs ou lectures métagénomiques) reposent à la fois sur la disponibilité des génomes ayant une qualité de séquençage et d'assemblage suffisante et sur l'utilisation de banques de données répertoriant les voies métaboliques et leurs fonctions. L'une des difficultés actuelles dans la détermination du potentiel fonctionnel d'un organisme est de pouvoir disposer de données fiables qui évoluent en fonction des derniers développements technologiques et de l'avancée des connaissances.

Problématiques et déroulement du travail expérimental

Pour piloter et contrôler les services écosystémiques rendus par les écosystèmes microbiens il convient d'identifier les espèces en présence dans l'écosystème, et de disposer d'une connaissance approfondie de leurs fonctions au sein de l'écosystème.

Comme exposé précédemment, l'identification par culture des procaryotes reste difficile notamment pour les espèces dont les besoins en nutriments ou conditions environnementales particulières sont largement inconnus ou non réplicables en laboratoire. L'identification est alors basée principalement sur le séquençage de l'ADN, soit partiellement par approche amplicon d'un gène marqueur de biodiversité, soit totalement par approche métagénomique (Partie A.2, page 42). Les avancées en technologie de séquençage et le raffinement des pipelines d'analyses bioinformatiques rendent l'approche métagénomique de plus en plus accessible à la communauté scientifique. L'une des méthodes d'analyse de ces séquences métagénomiques aboutit à la reconstruction de génomes putatifs ou espèces métagénomiques (Partie A.2.2.2, page 46). Dans la première partie de notre étude expérimentale nous présentons la problématique de correction d'assignation taxonomique d'espèces métagénomiques en utilisant une approche par reconstruction d'un arbre phylogénétique d'une part et en utilisant un indice global de parenté génomique d'autre part. Ce travail a été initié au cours d'un séjour dans l'équipe du Professeur Nicola Segata de l'université de Trente en Italie, à la faveur d'un financement INRA (DARESE, Agreenium) et de l'Université Paul Sabatier dans le cadre de la labellisation Agreenium de notre travail de thèse.

Afin de mieux appréhender le rôle fonctionnel des espèces présentes dans un écosystème, il est nécessaire d'accéder à une base de données fonctionnelles de haute qualité, en accès libre et interopérable. Au démarrage de notre travail de thèse cette base de données n'était pas disponible, c'est pourquoi nous avons créé la base de données MACADAM. Cette tâche qui constitue le coeur de notre travail de thèse est présentée dans le deuxième chapitre de

l'étude expérimentale sous la forme d'article publié dans Le Boulch et al. (2019) dans le journal DATABASE.

Deux exemples d'utilisation de MACADAM pour l'inférence du potentiel fonctionnel à partir de données taxonomiques sont présentées dans un troisième chapitre. Le premier exemple concerne l'inférence fonctionnelle de groupes taxonomiques bactériens dominants du microbiote cæcal de jeunes lapereaux dont les résultats ont été intégrés dans une publication, « Diversity and Co-occurrence Pattern Analysis of Cecal Microbiota Establishment at the Onset of Solid Feeding in Young Rabbits » Read et al. (2019). Dans un dernier chapitre de l'étude expérimentale nous confrontons les résultats de l'inférence fonctionnelle obtenus d'une part à partir de données métagénomiques sans assemblage et d'autre part à partir de données taxonomiques en utilisant MACADAM sur une communauté artificielle de onze espèces bactériennes.

Chapitre B

Études expérimentales

B.1 De l'affiliation taxonomique d'espèces métagénomiques au reclassement d'une espèce bactérienne

B.1.1 Contexte

Comme évoqué dans la partie introductive, le développement de la taxonomie procaryote a longtemps été limité par la capacité des microbiologistes à cultiver les procaryotes. En 1995, les deux premiers génomes complets de bactéries ont été séquencés (Fleischmann et al. 1995 ; Fraser et al. 1995) et ont marqué le début d'une augmentation drastique du nombre de génomes accessibles dans les banques de données (Land et al. 2015).

Désormais, l'évolution des techniques de séquençage permet le séquençage massif de l'ADN des procaryotes présents dans différents écosystèmes (approche métagénomique) et ne pouvant être cultivés. Parallèlement, les raffinements des méthodes bioinformatiques améliorent la qualité d'assemblage des séquences produites. Ces progrès technologiques conduisent actuellement à un afflux de production de nouveaux génomes souvent putatifs qu'il est nécessaire de classer sans autres connaissances phénotypiques. A ce contexte s'ajoute l'absence de consensus sur l'unité taxonomique de base qu'est l'espèce bactérienne que nous avons abordée dans la partie A.1.2, page 8, et les différences d'affiliations taxonomiques des procaryotes selon la base de données considérée (Partie A.1.7, page 32).

B.1.2 Problématique

La prise en compte des connaissances sur la diversité procaryotique des écosystèmes issus des approches métagénomiques est donc un réel challenge. Est-il possible d'identifier et de positionner dans la classification les organismes procaryotes dont les génomes sont reconstruits à partir de séquences métagénomiques ?

B.1.3 Objectifs

Pour apporter un début de réponse à cette question, nous baserons notre travail sur l'étude de deux genres bactériens, le genre *Blautia* et le genre *Ruminococcus*. L'objectif de ce travail est : (i) d'établir un état des lieux de la classification de ces genres, (ii) de comparer les similitudes et les différences entre ces deux genres grâce à la reconstruction d'un arbre phylogénétique et au calcul d'un indice de parenté génomique sur un jeu de données métagénomiques, (iii) de proposer un consensus taxonomique.

B.1.4 Matériels et méthodes

B.1.4.1 Description du jeu de données métagénomiques

Les données utilisées pour cette étude sont issues de Pasolli et al. (2019). Brièvement, les données métagénomiques sont obtenues à partir de prélèvements d'écosystèmes microbiens hébergés sur ou dans différentes parties du corps humain au sein d'une population occidentalisée ou non-occidentalisée (Figure B.I). Les lectures métagénomiques ont été assemblées en contigs et rassemblé en groupes (*bins*) via *binning* (Partie A.2.2.2, page 46). Les auteurs considèrent ces bins comme des « génomes présumés ». Ces génomes présumés sont réunis dans de nouveaux groupes au seuil de 95% de similarité génétique (SGB : Species-level Genome Bins). L'affiliation taxonomique de ces SGBs se fait par l'addition de génomes procaryotes issus de GenBank respectant le seuil de 95% de similarité. Deux types de groupes de génomes sont obtenus : (i) les kSGBs (pour known SGBs, n = 1 134), rassemblant des génomes reconstruits ainsi qu'un ou plusieurs génomes de GenBank. Ces groupes possèdent une affiliation taxonomique. (ii) Les uSGBs (pour unknown SGBs, n = 3 796), rassemblant des génomes reconstruits mais aucun génome de GenBank. Ces groupes ne sont donc pas affiliés taxonomiquement.

L'ensemble des kSGBs et uSGBs sont ensuite analysés phylogénétiquement et fonctionnellement. Les résultats de cette étude mettent en avant de nombreuses espèces microbiennes encore non-nommées et non cultivées ainsi qu'une prévalence des uSGBs chez les populations non-occidentalisées. Les

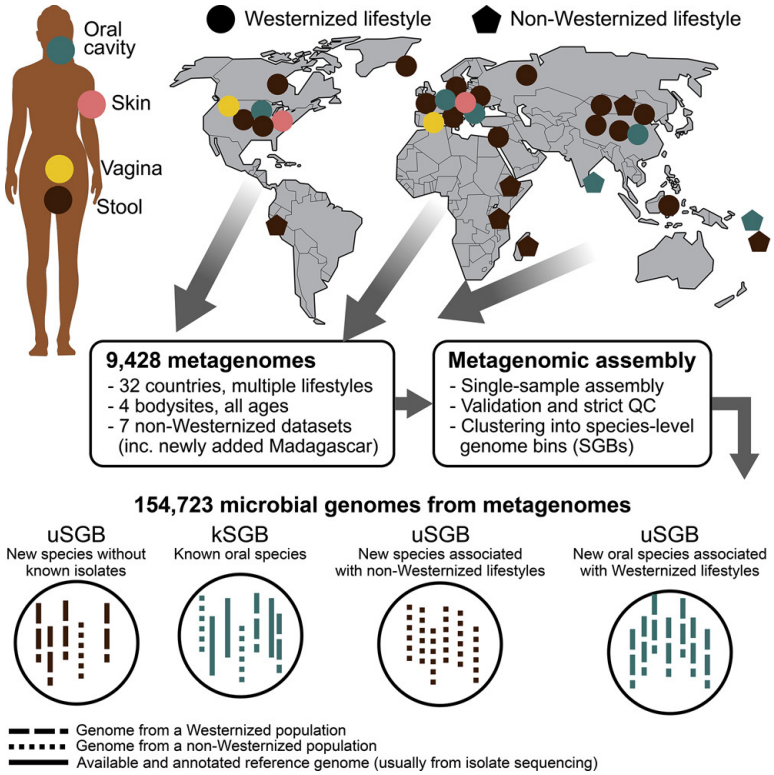


FIGURE B.I – Résumé graphique de l'étude PASOLLI et al. (2019). Ce schéma indique la nature des prélèvements et leurs origines géographiques. Ceux-ci sont séquencés afin d'obtenir des métagénomes qui sont ensuite assemblés afin de former des génomes reconstruits pouvant être associés, ou non, à des espèces connues.

Source: PASOLLI et al. (2019)

analyses fonctionnelles ont permis d'identifier des gènes du métabolisme du tryptophane présents uniquement chez les populations non-occidentalisées.

Dans notre étude, nous nous sommes intéressés aux kSGBs affiliés au genre *Blautia* et au genre *Ruminococcus*. Nous avons analysé 43 kSGBs représentant un total de 10 375 génomes reconstruits (Tableau B.1). Chaque kSGB contenant entre 1 et 1 925 génomes reconstruits, et chacune associée à 1 ou 16 génomes de GenBank, pour un total de 117 génomes GenBank. Notons que 9 de ces kSGBs portent l'affiliation taxonomique « *Bacteria* ; *Firmicutes* ; *Clostridia* ; *Clostridiales* ; *Lachnospiraceae* ; *Blautia* ; *Ruminococcus_sp* » qui est imprécise car *Ruminococcus_sp* ne devrait pas appartenir au genre *Blautia*.

kSGB ID	G. reconstruits	G. GenBank	ID alternatif	Taxonomie du génome le plus proche*
4184	104	2	4	Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Ruminococcus_sp
4208	17	1		Firmicutes ;Clostridia ;Clostridiales ;Ruminococcaceae ;Ruminococcus ;Ruminococcus_sCAG_624
4247	272	2	3	Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Ruminococcus_sp
4260	69	1		Firmicutes ;Clostridia ;Clostridiales ;Ruminococcaceae ;Ruminococcus ;Ruminococcus_sCAG_579
4262	1512	5	2	Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Ruminococcus_sp
4280	224	3	6	Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Ruminococcus_sp
4283	1	1		Firmicutes ;Clostridia ;Clostridiales ;Ruminococcaceae ;Ruminococcus ;Ruminococcus_bromii
4285	1925	5		Firmicutes ;Clostridia ;Clostridiales ;Ruminococcaceae ;Ruminococcus ;Ruminococcus_bromii
4367	868	1		Firmicutes ;Clostridia ;Clostridiales ;Ruminococcaceae ;Ruminococcus ;Ruminococcus_sCAG_177
4372	49	1		Firmicutes ;Clostridia ;Clostridiales ;Ruminococcaceae ;Ruminococcus ;Ruminococcus_sCAG_563
4373	140	1		Firmicutes ;Clostridia ;Clostridiales ;Ruminococcaceae ;Ruminococcus ;Ruminococcus_sCAG_488
4418	42	2		Firmicutes ;Clostridia ;Clostridiales ;Ruminococcaceae ;Ruminococcus ;Ruminococcus_champanellensis
4420	66	1		Firmicutes ;Clostridia ;Clostridiales ;Ruminococcaceae ;Ruminococcus ;Ruminococcus_sCAG_330
4422	233	1		Firmicutes ;Clostridia ;Clostridiales ;Ruminococcaceae ;Ruminococcus ;Ruminococcus_callidus
4423	61	1		Firmicutes ;Clostridia ;Clostridiales ;Ruminococcaceae ;Ruminococcus ;Ruminococcus_sCAG_403
4425	342	1		Firmicutes ;Clostridia ;Clostridiales ;Ruminococcaceae ;Ruminococcus ;Ruminococcus_sCAG_254
4552	19	2		Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Ruminococcus_sp
4557	263	2		Firmicutes ;Clostridia ;Clostridiales ;Ruminococcaceae ;Ruminococcus ;Ruminococcus_lactaris
4564	305	5		Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Ruminococcus_torques
4584	477	16		Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Ruminococcus_gnavus
4608	775	7	7	Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Ruminococcus_torques
4617	51	1		Firmicutes ;Clostridia ;Clostridiales ;Ruminococcaceae ;Ruminococcus ;Ruminococcus_sDSM_100440
4677	13	4		Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Blautia_hydrogenotrophica
4794	34	4		Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Blautia_hansenii
4800	3	2		Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Blautia_sAn46
4804	5	1		Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Blautia_sp
4810	58	1		Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Blautia_sCAG_237
4811	62	5		Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Blautia_obeum
4814	2	1		Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Blautia_obeum
4815	12	1	9	Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Ruminococcus_sp
4816	34	2		Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Blautia_sp
4820	192	5		Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Blautia_sp
4826	391	5		Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Blautia_sp
4828	133	2		Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Blautia_sp
4829	30	2		Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Blautia_sp
4844	1007	7		Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Blautia_obeum
4862	83	1		Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Blautia_sCAG_257
4868	113	2		Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Blautia_sp
4871	256	3	8	Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Ruminococcus_sp
5854	16	4	1	Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Ruminococcus_sp
14237	75	1		Firmicutes ;Clostridia ;Clostridiales ;Ruminococcaceae ;Ruminococcus ;Ruminococcus_sCAG_724
14243	40	1		Firmicutes ;Clostridia ;Clostridiales ;Ruminococcaceae ;Ruminococcus ;Ruminococcus_sCAG_382
14890	1	1		Firmicutes ;Clostridia ;Clostridiales ;Lachnospiraceae ;Blautia ;Ruminococcus_sp

Tableau B.1 – Description des 43 kSGBs appartenant aux genres *Blautia* et *Ruminococcus*. Chaque kSGB est décrite par son identifiant (ID) donné dans (Pasolli et al. 2019), son nombre de génomes reconstruits, les génomes GenBank qui lui sont rattachés ainsi que la taxonomie du génome GenBank présentant le plus de similarité avec les génomes reconstruits du kSGB. Un identifiant alternatif est donné aux 9 kSGBs présentant une affiliation erronée telle que décrite dans le texte. * la taxonomie est inférée à partir de l'étiquette taxonomique associée au génome GenBank. Elle est ensuite nettoyée de tous les mots ambigus dans le nom d'espèce.

B.1.4.2 Etat des lieux de la classification des genres *Ruminococcus* et *Blautia*

Ruminococcus Le genre *Ruminococcus* a été mis en évidence en 1948 (Sijpesteijn 1949) et validé en 1980 lors de la première publication de *Approved Lists of Bacterial Names* (Skerman, McGowan et Sneath 1989). *Ruminococcus flavefaciens* est la souche type du genre *Ruminococcus*. Il est le genre type de la famille des *Ruminococcaceae* qui a été mis en évidence par la même occasion et dont le nom a été validé en 2010 (Euzéby 2010 ; Rainey 2009) bien que celui-ci ait été effectif auparavant. La classification abrégée proposée par le NCBI Taxonomy pour le genre *Ruminococcus* est « *Bacteria ; Firmicutes ; Clostridia ; Clostridiales ; Ruminococcaceae* ».

LSPN	**NCBI Taxonomy**
Ruminococcus albus	*Ruminococcus albus*
	Ruminococcus bicirculans
Ruminococcus bromii	*Ruminococcus bromii*
Ruminococcus callidus	*Ruminococcus callidus*
Ruminococcus champanellensis	*Ruminococcus champanellensis*
Ruminococcus faecis	*Ruminococcus faecis*
Ruminococcus flavefaciens	*Ruminococcus flavefaciens*
Ruminococcus gauvreauii	*Ruminococcus gauvreauii*
Ruminococcus gnavus	
Ruminococcus lactaris	*Ruminococcus lactaris*
Ruminococcus torques	

Tableau B.2 – Les espèces présentes dans le genre *Ruminococcus* d'après List of Prokaryotic names with Standing in Nomenclature (LPSN) et NCBI Taxonomy.

Les *Ruminococci* possèdent des caractéristiques communes : les cellules sont de forme coccoïde, certaines pouvant être motiles via un ou des flagelles. Elles sont anaérobiques strictes, chimioorganotrophiques (leur énergie provient de composés organiques carbonés) et leur température optimale de croissance est de 37-42°C. Elles sont isolées à partir de contenus ruminaux ou de contenus de gros intestins et de cæcum de plusieurs animaux et de l'Humain. Au moment de l'écriture de cette thèse, le genre *Ruminococcus* contient 10 espèces selon LPSN et 9 espèces selon NCBI Taxonomy (Tableau B.2).

Blautia Le genre *Blautia* a été décrit en 2008 (Liu et al. 2008). L'espèce type est *Blautia coccoides*. Il fait partie de la famille des *Lachnospiraceae* (Lawson et Finegold 2015) bien qu'il ne soit pas affilié à une famille dans le LPSN. La classification abrégée proposée par le NCBI Taxonomy pour le genre *Blautia* est « *Bacteria ; Firmicutes ; Clostridia ; Clostridiales ; Lachnospiraceae* ».

Le genre *Blautia* contient des organismes présentant les caractéristiques communes suivantes : les cellules sont de forme coccoïde ou ovale. Elles sont anaérobies obligatoires et présentent un métabolisme de type chimioorganotrophiques. Elles ont été isolées a partir de fèces humains ou animaux (Durand et al. 2017). A ce jour, le genre *Blautia* est composé de 13 espèces selon le LPSN et de 19 selon le NCBI Taxonomy (Tableau B.3).

LPSN	**NCBI Taxonomy**
Blautia caecimuris	*Blautia caecimuris*
Blautia coccoides	*Blautia coccoides*
Blautia faecis	*Blautia faecis*
Blautia glucerasea	*Blautia glucerasea*
Blautia hansenii	*Blautia hansenii*
Blautia hominis	*Blautia hominis*
Blautia hydrogenotrophica	*Blautia hydrogenotrophica*
Blautia luti	*Blautia luti*
	Blautia marasmi
	Blautia massiliensis
Blautia obeum	*Blautia obeum*
	Blautia phocaeensis
Blautia producta	*Blautia producta*
	Blautia provencensis
Blautia schinkii	*Blautia schinkii*
Blautia stercoris	*Blautia stercoris*
Blautia wexlerae	*Blautia wexlerae*
	*[Ruminococcus] gnavus**
	*[Ruminococcus] torques**

Tableau B.3 – Les espèces présentes dans le genre *Blautia* d'après LPSN et NCBI Taxonomy. * : ces deux espèces sont en attente de transfert dans un autre genre. Leurs noms n'ont pas encore été publiés dans la littérature appropriée.

Deux genres proches : Les genres *Ruminococcus* et *Blautia* ont une histoire commune. Le genre *Ruminococcus*, décrit en premier, est un groupe polyphylétique (groupe d'organismes n'ayant pas d'ancêtre commun direct) car il est composé d'espèces provenant de deux familles : la famille des *Ruminococcaceae* et des *Lachnospiraceae* (Rajilić-Stojanović et Vos 2014). Liu et al. (2008) ont donc proposé un nouveau genre de la famille des Lachnospiraceae : *Blautia*. Lors de la description de ce genre, des espèces du genre *Ruminococcus* ont été transférées dans ce genre (Lawson et Finegold 2015 ; Togo et al. 2018). On retrouve ces changements dans les tableaux B.2 et B.3 : (i) selon le LPSN, *Ruminococcus gnavus* et *Ruminococcus torques* appartiennent toujours au genre *Ruminococcus* mais sont en cours de transfert vers un autre genre selon le NCBI Taxonomy. Selon Liu et al. (2008), ces deux espèces n'appartiendraient ni au genre *Blautia* ni au genre *Ruminococcus*. Togo et al. (2018) isolent une espèce bactérienne et définissent un nouveau genre de la famille Lachnospiraceae qu'ils nomment *Mediterraneibacter* gen. nov. . Il est proposé que *Ruminococcus gnavus* et *Ruminococcus torques* rejoingne ce nouveau genre. (ii) Les espèces *Ruminococcus faecis* et *Ruminococcus lactaris* ont été également identifiées comme n'appartenant pas au genre *Ruminococcus* mais au genre *Mediterraneibacter* par Togo et al. (2018). Ces changements n'ont pas été répercutés à ce jour ni sur le NCBI Taxonomy, ni dans la liste LPSN. (iii) L'espèce *Ruminococcus bicirculans* n'a pas été publiée dans la littérature appropriée et n'a pas été déposée dans les deux bases de données d'espèces bactériennes nécessaires à sa déclaration dans la nomenclature bactérienne. (iv) De nouvelles espèces du genre *Blautia* ont été isolées et ne sont pas encore présentes dans le NCBI Taxonomy (Shin et al. 2018 ; Paek et al. 2019). L'ensemble de ces changements taxonomiques sont résumés dans le tableau B.4. Les genres *Blautia* et *Ruminococcus* nécessitent donc encore une attention particulière afin de clarifier les espèces présentes dans chacun d'entre eux.

Nom d'espèce initial	Raison de la différence de nomenclature	Référence
R. bicirculans	Non dépôt dans deux collections	[1]
R. faecis	En cours de transfert vers le genre *Mediterraneibacter*	[2]
R. gnavus	En cours de transfert vers le genre *Mediterraneibacter*	[2]
R. lactaris	En cours de transfert vers le genre *Mediterraneibacter*	[2]
R. torques	En cours de transfert vers le genre *Mediterraneibacter*	[2]
B. argi	Publié et validé. En cours d'intégration dans LPSN	[3]
B. hominis	Publié et validé. En cours d'intégration dans LPSN	[4]
B. marasmi	Non dépôt dans deux collections	[5]
B. massiliensis	Non validé	[6]
B. phocaeensis	Non dépôt dans deux collections	[7]
B. provencensis	Non dépôt dans deux collections	[8]
M. massiliensis	Publié et validé. En cours d'intégration dans LPSN	[2]

Tableau B.4 – Raison des différences de nomenclatures entre le LPSN et le NCBI Taxonomy dans les genres *Blautia, Ruminococcus* et *Mediterraneibacter*. Légende des références [1] Wegmann et al. (2014) , [2] Togo et al. (2018) et Oren et Garrity (2019a) , [3] Paek et al. (2019) et Oren et Garrity (2019b) , [4] Shin et al. (2018) et Oren et Garrity (2018b) , [5] Pham et al. (2017a) , [6] Durand et al. (2017) , [7] Traore et al. (2017) , [8] Pham et al. (2017b).

B.1.4.3 PhyloPhlAn : pipeline de classification phylogénétique des génomes reconstruits et des génomes RefSeq

Afin de pouvoir clarifier la position des kSGBs dans les genres *Blautia* et *Ruminococcus* nous avons utilisé PhyloPhlAn. L'analyse portera sur les 5 génomes reconstruits de chaque kSGB auxquels ont été ajoutés l'intégralité des génomes RefSeq appartenant aux genres *Blautia* (n = 50) et *Ruminococcus* (n = 89). PhyloPhlAn est un pipeline d'outils permettant de réaliser une classification phylogénétique des génomes et génomes reconstruits à partir de données métagénomiques (Segata et al. 2013). Nous avons utilisé la deuxième version de PhyloPhlAn, non publiée à l'heure de l'écriture de ce manuscrit, qui optimise la vitesse d'exécution et permet de régler plus finement les paramètres des outils présents dans le pipeline. La documentation de la deuxième version de PhyloPhlAn est disponible à l'adresse suivante : `https://bitbucket.org/nsegata/phylophlan/wiki/phylophlan2`. PhyloPhlAn inclut une base de données de 400 protéines qui sont des marqueurs phylogénétiques. USEARCH (Edgar 2010) est utilisé pour identifier ces marqueurs protéiques dans les génomes d'intérêts par alignement traduit. Ces alignements servent ensuite de fichiers d'entrée à FastTree (Price, Dehal et Arkin 2010) qui permet de reconstruire un premier arbre phylogénétique par maximum de vraisemblance. Une fois cet arbre obtenu, RAxML (Stamatakis 2014) est utilisé sur l'arbre obtenu par FastTree afin d'augmenter la précision de sa topologie et de minimiser le score de parcimonie de l'arbre (Liu, Linder et Warnow 2011 ; Segata et al. 2013). Au vue de notre jeu de données, nous utilisons PhyloPhlAn avec les options « accurate » et « diversity low » car les génomes reconstruits composant notre jeu de données sont issus de deux genres proches. Mais ces deux options demandent un temps de calcul important si les calculs sont demandés sur l'intégralité des génomes reconstruits. Sachant que les génomes reconstruits sont regroupés en fonction de leur similarité génétique (kSGB), un script est développé afin de tirer aléatoirement un maximum de 5 génomes reconstruits dans chaque groupe kSGB permettant de diminuer le temps de calcul de l'arbre phylogénétique. Afin de pouvoir enraciner l'arbre, nous avons intégré dans nos données le génome de *Clostridioides difficile* 630 (anciennement connue sous le nom de *Clostridium difficile* 630 (Hall et O'toole 1935 ; Prévot 1938 ; Holdeman, Cato et Moore 1977 ; Skerman, McGowan et Sneath 1980) mais celle-ci a été sujette d'un reclassement (Lawson et al. 2016 ; Oren et Garrity 2016) provenant de RefSeq (numéro d'accession : GCF_000009205.2). *Clostridioides difficile 630* appartient, comme les genres *Blautia* et *Ruminococcus*, à l'ordre des *Clostridiales*.

B.1.4.4 Calcul du score ANI

Nous utiliserons le score ANI (Partie A.1.5.2, page 24) comme métrique pour établir l'appartenance ou non de différents génomes à une même espèce. Pour cela nous avons utilisé un module Python pyani (`https://widdowquinn.github.io/pyani/`). Il existe quatre méthodes (ANIb, ANIm, FastANI et OrthoANI) pour calculer le score ANI entre deux génomes. Nous avons utilisé ici l'ANIm reposant sur MUMmer (Kurtz et al. 2004). Ce choix s'explique par la proximité des génomes étudiés, ANIm étant plus robuste dans ce cas (Richter et Rosselló-Móra 2009) et par sa rapidité par rapport aux autres méthodes (Yoon et al. 2017b). L'ANI s'exprime en pourcentage. Quand il est supérieur à 96%, l'ANI indique que deux génomes appartiennent à la même espèce. Lorsqu'il est compris entre 93% et 96%, alors il existe une incertitude quant à l'appartenance à la même espèce des deux génomes et il est nécessaire d'utiliser d'autres techniques telles que celles basées sur le gène de l'ARNr 16S afin de confirmer ou d'infirmer l'hypothèse. Quand l'ANI est inférieur à 93% alors les deux génomes ne font pas parti de la même espèce.

B.1.5 Résultats et Discussion

B.1.5.1 Analyse taxonomique des groupes d'espèces métagénomiques à partir de l'arbre phylogénétique

Notre première étape est de placer les génomes reconstruits (n = 215) issus des 43 kSGBs sur un arbre phylogénétique reconstruit par PhyloPhlAn. Les résultats sont donnés dans la figure B.II, page 89.

Le génome de *Clostridioides difficile*, utilisé pour enraciner l'arbre est symbolisé par le A pour Ancestor. On distingue quatre branches majeures dans l'arbre dont trois regroupent les espèces dont les génomes sont issus de RefSeq (Tableau B.5) : (i) la branche rose contient l'ensemble des espèces de *Ruminococcus* que le NCBI Taxonomy considère comme appartenant au genre *Ruminococcus*, (ii) la branche bleue contient l'ensemble des espèces de *Ruminococcus* en cours de transfert vers le genre *Mediterraneibacter* (Togo et al. 2018) ainsi que l'espèce type de ce genre : *Mediterraneibacter massiliensis* (Figure B.II **B**) et (iii) la branche violette contient l'ensemble des génomes des espèces du genre *Blautia* ainsi qu'une espèce du genre *Ruminococcus* : *R. gauvreauii* (Figure B.II **C**).

Les parties bleue et violette sont quant à elles plus proches entre elles que de la partie rose. Ainsi, selon cet arbre phylogénétique, les espèces affiliées au genre *Blautia* sont plus proches du genre *Mediterraneibacter* que du genre *Ruminococccus* en lien avec leur appartenance à la famille des *Lachnospiraceae* (*Blautia* et *Mediterraneibacter*) et *Ruminoccocaceae* (*Ruminococccus*).

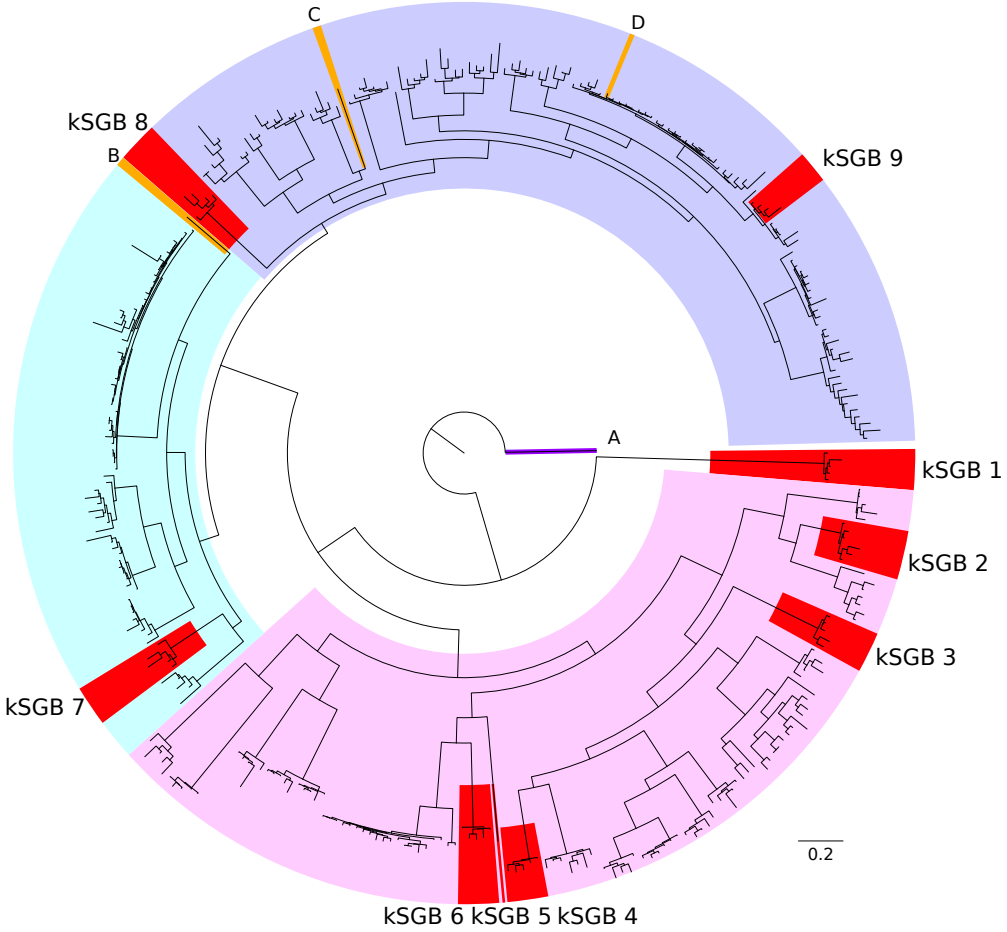


FIGURE B.II – Arbre phylogénétique se basant sur les cinq génomes reconstruits sélectionnés aléatoirement de chacun des 43 kSGBs identifiés comme appartenant au genre *Blautia* et *Ruminococcus* ainsi que les 50 et 89 génomes RefSeq pour les genres *Blautia* et *Ruminococcus*, respectivement. Les kSGBs numérotés de 1 à 9 présentent une affiliation taxonomique erronée. **A** : *Clostridioides difficile* utilisé pour l'enracinement de l'arbre. **B** : *Mediterraneibacter massiliensis*. **C** : génome RefSeq de la souche type de l'espèce *Ruminococcus gauvreauii*. **D** : génome RefSeq de la souche type de l'espèce *Blautia obeum*.

Branche rose	**Branche bleue**	**Branche violette**
R. albus	*M. massiliensis*	*B. coccoides*
R. bicirculans	*R. faecis*	*B. hansenii*
R. bromii	*R. gnavus*	*B. hominis*
R. callidus	*R. lactaris*	*B. hydrogenotrophica*
R. champanellensis	*R. torques*	*B. marasmi*
R. flavefaciens		*B. massiliensis*
		B. obeum
		B. producta
		B. schinkii
		B. wexlerae
		R. gauvreauii

Tableau B.5 – Position des génomes issus de RefSeq dans l'arbre phylogénétique de la figure B.II, page 89.

Le groupe kSGB 1 n'est inclus dans aucune des trois branches. Les génomes formant ce groupe ne feraient donc pas partie du genre *Ruminococcus*, *Blautia* ou *Mediterraneibacter*.

Le génome de référence associé au groupe kSGB 1 lors de la constitution est issu de GenBank (GCA_900066605). La taxonomie de ce génome de référence est selon le NCBI : *Bacteria*, *Firmicutes*, *Clostridia*, *Clostridiales*, *Ruminococcaceae*, *Ruminococcus*, environmental samples. Ce génome provient d'un organisme non cultivable, issu du séquençage d'un environnement. Il est donc fortement possible que ce génome ait été déposé dans GenBank avec une erreur dans sa taxonomie.

Lors de sa constitution, le groupe kSGB 1 est également associé avec trois autres génomes de référence plus distants (Tableau B.1, page 82). Ces génomes sont les suivants : GCF_002163455, GCF_002287425 et GCA_900066485. Les deux premiers proviennent de RefSeq et sont affiliés au genre *Megasphaera* (« *Bacteria ; Firmicutes ; Negativicutes ; Veillonellales ; Veillonellaceae* ») tandis que le dernier provient de GenBank, il est issu d'un échantillonage environnemental et est affilié au genre *Megasphaera* également. L'ensemble de ces observations milite en faveur d'un rapprochement du groupe kSGB 1 au genre *Megasphaera* si l'on admet l'erreur de taxonomie lors du dépôt du génome GCA_900066605.

Les kSGBs présentant une affiliation taxonomique erronée et notée kSGBs 2, 3, 4, 5 et 6 semblent appartenir au genre *Ruminococcus* (partie rose de l'arbre). La partie de l'arbre regroupant les 5 génomes aléatoirement choisis du groupe kSGB 2 contient également le génome RefSeq de la souche type de

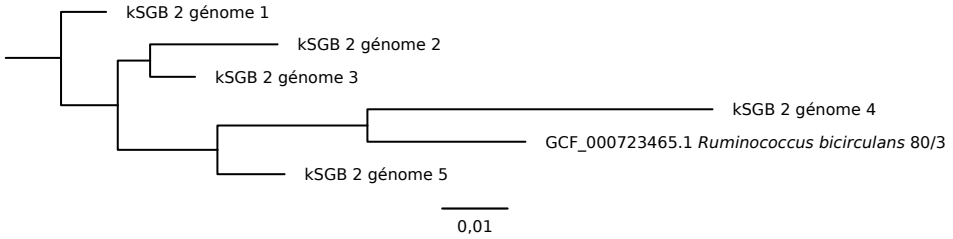


FIGURE B.III – Zoom sur la zone de l'arbre phylogénétique contenant les génomes reconstruits du groupe kSGB 2.

l'espèce *Ruminococcus bicirculans* 80/3 (GCF_000723465.1) (Figure B.III).

Les génomes reconstruits du groupe kSGB 7 bien qu'inclus dans la zone bleue et donc proches du genre *Mediterraneibacter*, occupent la position la plus éloignée du reste des bactéries de ce groupe.

De même, les génomes reconstruits du groupe kSGB 8 inclus dans la branche violette et donc proches du genre *Blautia* présentent une situation distante du reste du groupe et aucun génome de référence n'est présent à sa proximité.

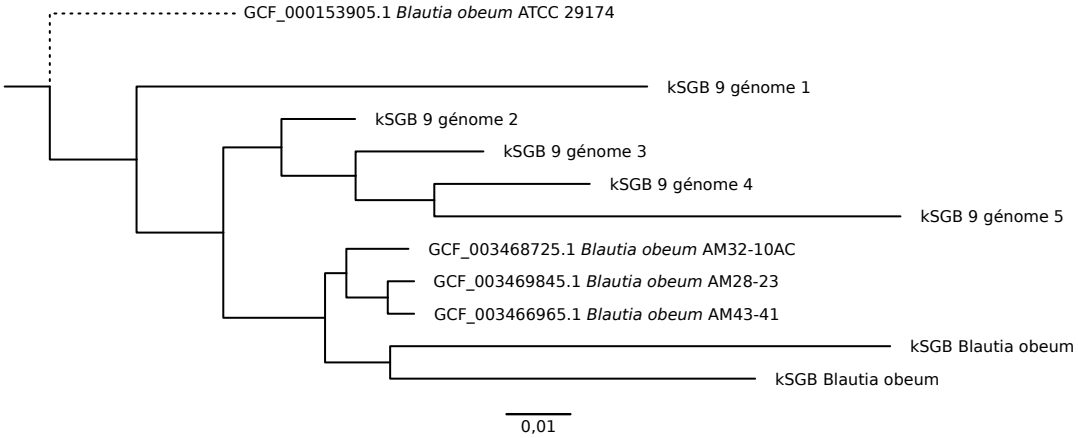


FIGURE B.IV – Zoom sur la zone du groupe kSGB 9. Les génomes notés « kSGB Blautia obeum » font partie d'une autre kSGB affiliée à *Blautia obeum*. Blautia obeum ATCC 29174 est le génome de la souche type de l'espèce *Blautia obeum*. Le point de divergence avec le génome de la souche type est indiqué en pointillés.

Les génomes du groupe kSGB 9 sont à proximité de certains génomes RefSeq de l'espèce *Blautia obeum* (Figure B.IV).Toutefois la souche type de l'espèce *Blautia obeum* n'est pas positionnée à proximité des entités du groupe kSGB 9 et des 3 souches (AM32-10AC, AM28-23, AM43-41) de l'es-

pèce (Figure B.II **D**). La question de la validité de l'affiliation taxonomique à l'espèce *B. obeum* se pose donc pour l'ensemble des entités présentes sur la branche de l'arbre phylogénétique présentée à la figure B.IV.

Concernant la position du génome RefSeq de la souche type de l'espèce *Ruminococcus gauvreauii* (Figure B.II **C**) se positionne dans l'arbre dans la branche violette où se situe les génomes RefSeq du genre *Blautia*. Togo et al. (2018) évoquait la possibilité d'un reclassement de l'espèce *Ruminococcus gauvreauii* du genre *Ruminococcus* au genre *Blautia*. Nos résultats semble indiquer que en effet, l'espèce *Ruminococcus gauvreauii* n'appartient pas au genre *Ruminococcus* mais semblerait appartenir au genre *Blautia*.

Les résultats de l'analyse de l'arbre phylogénétique obtenus sont présentés dans le tableau B.6.

kSGB	**Affiliation taxonomique par analyse de l'arbre phylogénétique**
kSGB 1	*Megasphaera* sp.
kSGB 2	*Ruminococcus bircirculans*
kSGB 3	*Ruminococcus* sp.
kSGB 4	*Ruminococcus* sp.
kSGB 5	*Ruminococcus* sp.
kSGB 6	*Ruminococcus* sp.
kSGB 7	proche *Mediterraneibacter*
kSGB 8	proche *Blautia*
kSGB 9	*Blautia* sp.

Tableau B.6 – Affiliations taxonomiques possibles des 9 kSGBs d'intérêt à la suite de l'analyse de l'arbre phylogénétique (Figure B.II, page 89).

B.1.5.2 Analyse taxonomique des groupes d'espèces métagénomiques à partir du calcul du score ANIm

Dans un deuxième temps, le score ANIm a été calculé entre les génomes RefSeq les génomes reconstruits des groupes kSGB 1, 2 et 9.

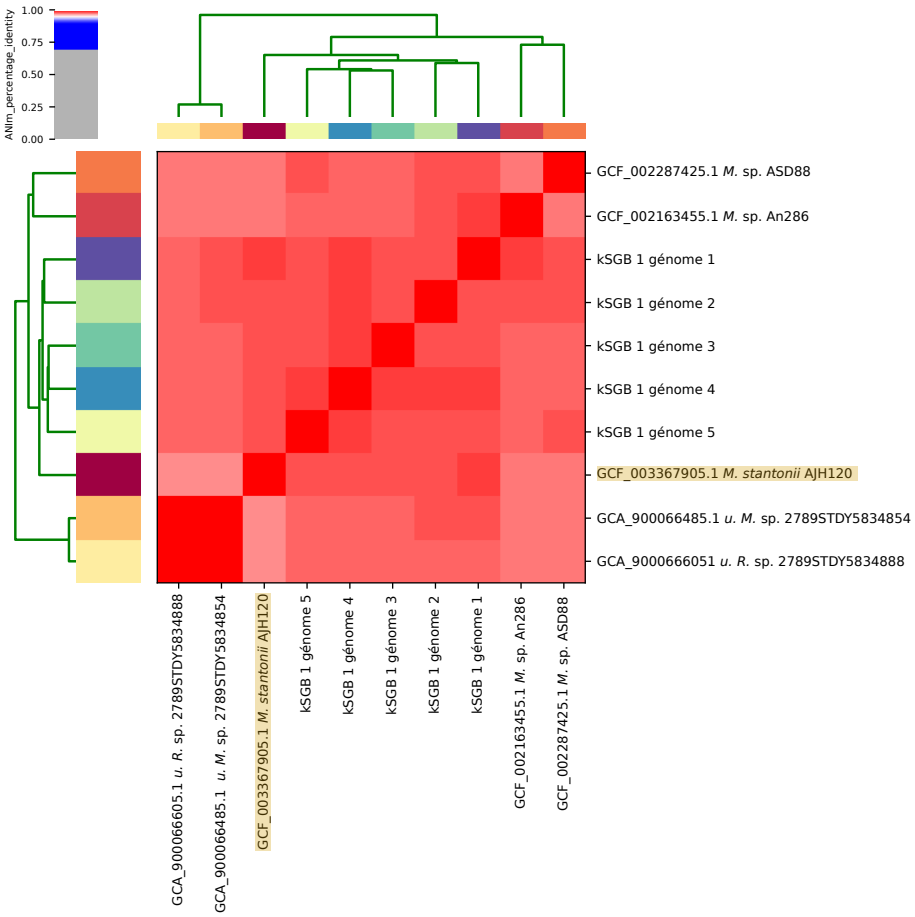


FIGURE B.V – Représentation de l'ANIm calculé entre les 5 génomes choisis aléatoirement dans le groupe kSGB 1 et les 4 génomes RefSeq et GenBank associés au groupe kSGB 1 ainsi que le génome de la souche type de l'espèce *Megasphaera stantonii* : *Megasphaera stantonii* AJH120.

Le score ANIm calculé entre les 5 génomes choisis aléatoirement du groupe kSGB 1, des 4 génomes de références associé au groupe kSGB 1 ainsi que le génome RefSeq de la souche type de l'espèce *Megasphaera stantonii* (Figure B.V), est supérieur à 97%. Cela semble signifier que le groupe kSGB 1 peut être affilié à l'espèce *Megasphaera stantonii*. Il en va de même pour le génome de référence associé au groupe kSGB 1 qui ne serait pas un *uncultured*

Ruminococcus sp. mais qui appartiendrait également à l'espèce *Megasphaera stantonii*.

Le score ANIm calculé entre 5 génomes reconstruits choisis aléatoirement dans le groupe kSGB 2 et *Ruminococccus bicirculans* est supérieur à 97% (Figure B.VI) soutenant ainsi l'hypothèse que le groupe kSGB 2 soit constitué d'organismes appartenant à *Ruminococcus bicirculans*.

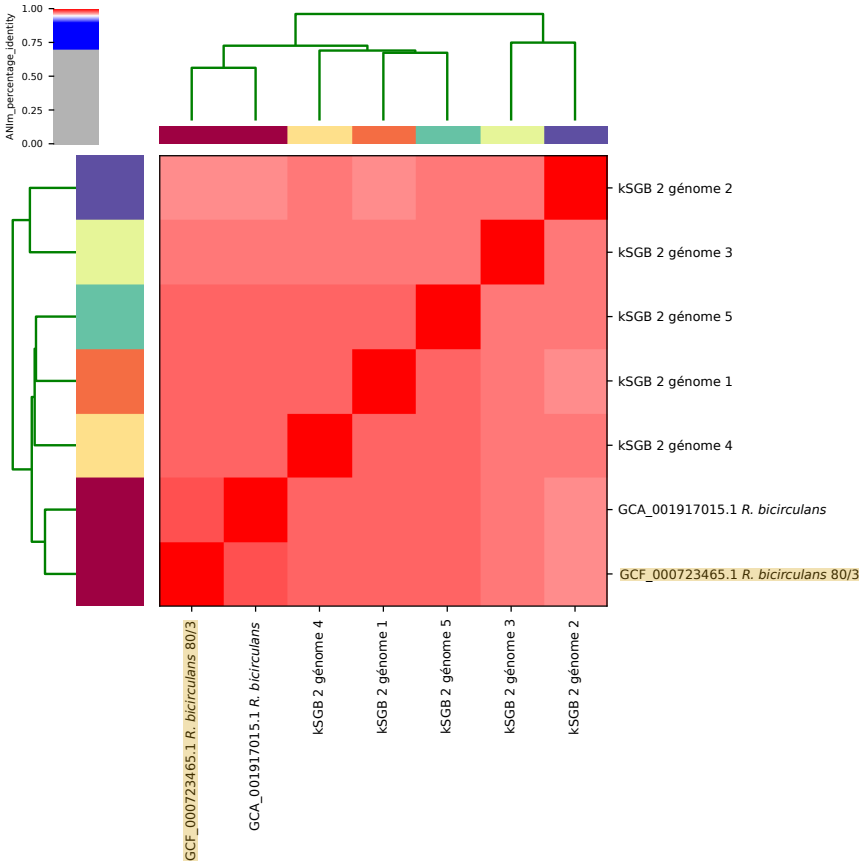


FIGURE B.VI – Représentation de l'ANIm calculé entre les 5 génomes choisis aléatoirement dans le groupe kSGB 2, les 2 génomes RefSeq et les 2 génomes GenBank de *Ruminococcus bicirculans*. Ruminococcus bicirculans 80/3 est la souche type de l'espèce.

En ce qui concerne le groupe kSGB 9, les résultats du calcul du score ANIm présentés dans la figure B.VII indiquent que les 5 génomes reconstruits du groupe kSGB 9 seraient représentatifs d'une même espèce (ANI $\geq$ 96%). Toutefois : (i) les scores ANIm obtenus entre les trois génomes de référence et les génomes 1, 2 et 4 du groupe kSGB 9 sont inférieurs à 96% et (ii) la souche type de l'espèce *Blautia obeum* présente quant à elle un score ANIm

éloigné de tous les génomes (ANIm <88%).

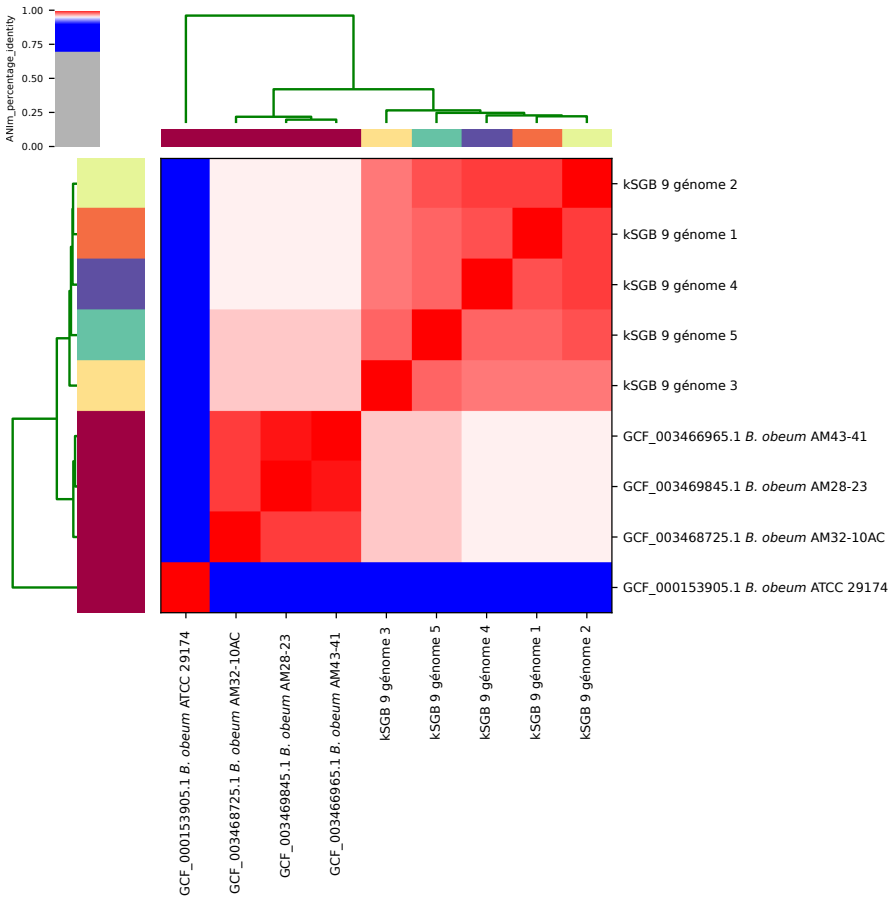


FIGURE B.VII – Représentation de l'ANIm calculé entre les 5 génomes choisis aléatoirement du groupe kSGB 9 et une sélection de génomes assemblés affiliés à *Blautia obeum* présent dans RefSeq. *Blautia obeum* ATCC 29174 est la souche type de l'espèce.

En conclusion, si l'on se réfère au score ANIm le groupe kSGB 9 constitue une espèce, proche de certaines souches dite *B. obeum*. Toutefois compte tenu de la distance avec la souche type de l'espèce *B. obeum*, ce groupe devrait être transféré dans une nouvelle espèce.

L'ensemble de ces résultats est récapitulé dans le tableau B.7. Il nous est possible de confirmer les affiliations taxonomiques données dans le tableau B.6 pour les kSGBs 1, 2 et 9. Le groupe kSGB 1 serait formé de génomes appartenant à l'espèce *Megasphaera stantonii* (MAKI et LOOFT 2018), espèce identifiée dans le cæcum des poules et productrice de butyrate. Il nous est également possible de proposer un reclassement des génomes RefSeq et

GenBank associés à cette kSGB. En ce qui concerne le groupe kSGB 2, le score ANI confirme les résultats de l'arbre à savoir son affiliation à l'espèce *Ruminococcus bicirculans*. Le groupe kSGB 9 semble elle appartenir à l'espèce *Blautia obeum* mais la position de la souche type de cette espèce semble contredire ce résultat.

kSGB	**Affiliation taxonomique par calcul de l'ANI**
kSGB 1	*Megasphaera stantonii*
KSGB 2	*Ruminococcus bircirculans*
KSGB 9	*Blautia* sp.

Tableau B.7 – Affiliation taxonomique possible des 3 kSGBs possédant des génomes RefSeq à proximité (Figure B.V, B.VI, et B.VII).

B.1.5.3 Calcul des scores ANIm des génomes RefSeq du genre *Blautia*

A la suite des précédents résultats nous décidons de calculer le score ANIm sur l'intégralité des génomes *Blautia obeum* assemblés présents dans RefSeq et GenBank (n = 34) afin d'affiner la position du groupe kSGB 9. Un seul des génomes appartenant à l'espèce *Blautia obeum* est présent dans GenBank. Les résultats sont présentés dans la figure B.VIII. On observe 4 groupes distincts de génomes affiliés à l'espèce *Blautia obeum* (notés A, B, C, et D). Entre eux, ces 4 groupes présentent un score ANIm inférieur ou égale à 90% . Ils représenteraient donc potentiellement 4 espèces bactériennes différentes. Le groupe D contenant le plus de génomes contient la souche type, *Blautia obeum* ATCC 29174, ainsi que le seul génome présent uniquement dans GenBank, *Blautia obeum* AM27-32LB.

Pour reclasser les trois groupes notés A, B et C de la figure B.VIII, n'incluant pas l'espèce type, nous rajoutons pour le calcul du score ANIm des génomes des espèces *Blautia wexlerae* (Liu et al. 2008) et de *Blautia massiliensis* (Durand et al. 2017) (souches type respectives : *Blautia wexlerae* DSM 19850 et *Blautia massiliensis* GD9 (Figure B.IX).

Concernant le groupe de génomes noté C, aucun génome de référence n'a pu être associé au seuil de 96%. Or, ce groupe contient les génomes RefSeq proches phylogénétiquement (Figure B.IV) des génomes reconstruits du groupe kSGB 9. Il n'est donc pas possible de préciser pour ce groupe kSGB 9 l'affiliation d'espèce dans le genre *Blautia*.

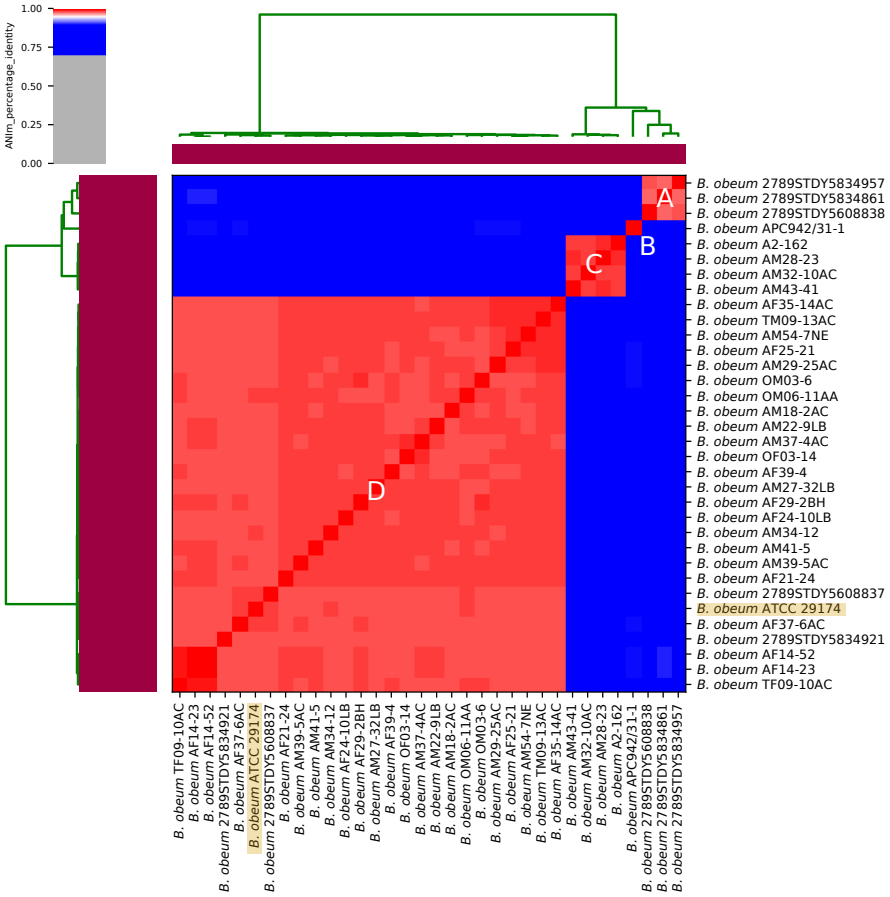


FIGURE B.VIII – Ensemble des génomes assemblés appartenant à l'espèce *Blautia obeum* présent dans RefSeq (n = 33) et GenBank (n = 1). *Blautia obeum* ATCC 29174 surlignée correspond à la souche type de l'espèce.

A l'inverse, le score ANIm positionne des génomes de souche type dans les groupes A et B. Ainsi *B. obeum* APC942/31-1 (groupe B) pourrait être reclassé à l'espèce *Blautia massiliensis* et les souches du groupe A comme appartenant à *Blautia wexlerae*.

Dans cette figure, les génomes des espèces *Blautia wexlerae* et *Blautia massiliensis* (souches type respectives : *Blautia wexlerae* DSM 19850 et *Blautia massiliensis* GD9). *B. obeum* APC942/31-1 semblent donc appartenir à l'espèce *Blautia massiliensis* tandis que trois souches de *B. obeum* (2789STDY5834861, 2789STDY5834957, 2789STDY5608838) semblent appartenir à l'espèce *Blautia wexlerae*. Il est à noter que les génomes de ces trois dernières souches ainsi que *B. obeum* APC942/31-1 sont présentes dans RefSeq sous forme de contig ou scaffold.

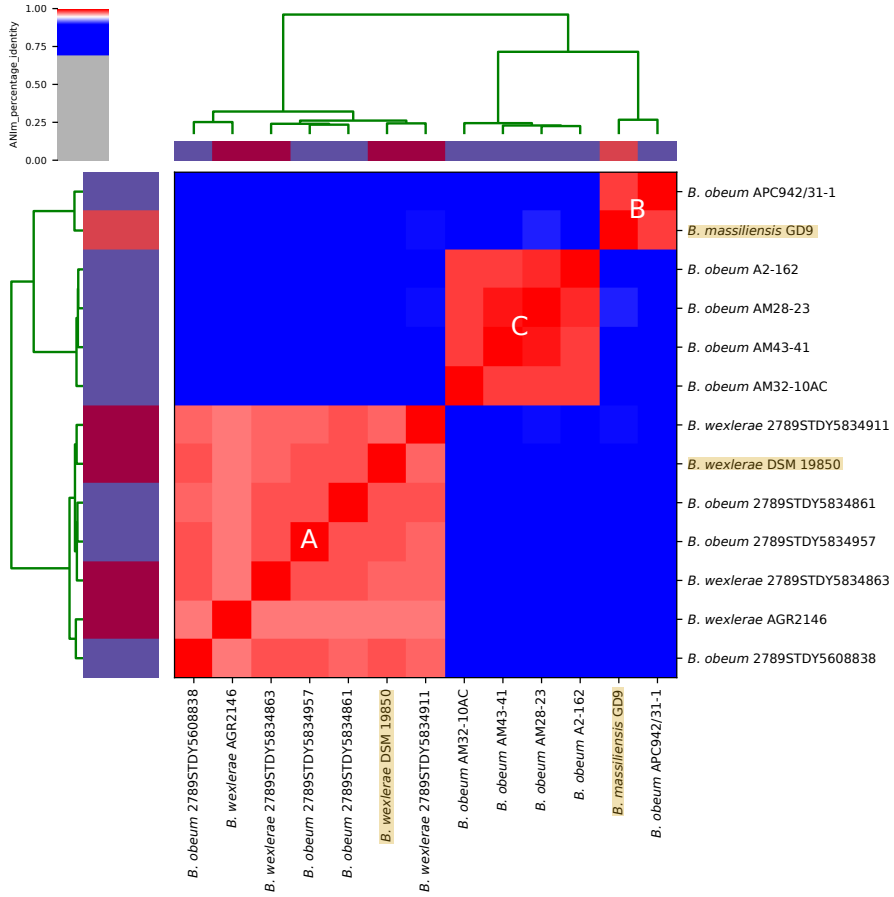


FIGURE B.IX – Ensemble des génomes assemblés appartenant aux espèces *Blautia obeum*, *Blautia wexlerae* (n = 4) et *Blautia massiliensis* (n = 1) présents dans RefSeq. Les génomes de *B. obeum* proviennent des trois groupes A, B et C de la figure B.VIII.

Les résultats du calcul du score ANI entre les génomes RefSeq des espèces *Blautia obeum*, *Blautia massiliensis* et *Blautia wexlerae* B.8 permettent le reclassement de certains génomes RefSeq. Ces résultats mettent en lumière la possibilité de proposer une nouvelle espèce formée par les génomes kSGB 9 et ceux du groupe C. Il est à noter que les quatre génomes des organismes du groupe C sont assemblés sous forme de contig, voire même de chromosome pour la souche A2-162 [1].

Afin de confirmer que les quatre organismes du groupe C n'appartiennent

1. Le génome de *Blautia obeum* A2-162 a été supprimé de RefSeq depuis ces résultats car l'assemblage contenait des erreurs de validation et n'incluait pas les gènes ribosomaux notamment. Il reste cependant toujours disponible dans Genbank, affilié à *Blautia obeum*.

pas à l'espèce *Blautia obeum*, il serait nécessaire de reconstruire un arbre phylogénétique basé sur de multiples marqueurs (Figueras et al. 2014). En outre, lors de la manipulation de ce groupe de génomes provenant de banques de données publiques, nos résultats montrent qu'il peut exister un doute dans l'affiliation taxonomique donnée ce qui a été relevé dans d'autres genres (Beaz-Hidalgo et al. 2015).

Groupe	Affiliation taxonomique	Raison
A	*Blautia wexlerae*	Présence de la souche type
B	*Blautia massiliensis*	Présence de la souche type
C	*Blautia* sp.	Pas de souche type
D	*Blautia obeum*	Présence de la souche type

Tableau B.8 – Résultats du calcul des scores ANI sur les génomes RefSeq des espèces *Blautia obeum*, *Blautia massiliensis* et *Blautia wexlerae* des groupes définis dans la figure B.VIII, page 97.

B.1.6 Conclusions

Dans cette partie nous avons repositionné dans la classification 9 kSGBs grâce à la reconstruction d'un arbre à partir de protéines marqueurs dans un premier temps puis avec un indice global de parenté génomique (ANI). Nous avons également mis en évidence des erreurs d'affiliation taxonomique pour des génomes provenant de GenBank, voire de RefSeq dans les espèces *Megasphaera stantonii* et *Blautia obeum*. L'examen plus approfondi de ces génomes montre qu'au moins quatre d'entre eux proviennent de l'étude de Browne et al. (2016). Ceci montre bien les limites de l'affiliation taxonomique des séquences déposées dans les banques de données génomiques. En effet lors du dépôt d'un génome dans GenBank aucune analyse postérieure au dépôt n'est effectuée par le personnel du NCBI. L'erreur proviendrait donc de la taxonomie associée lors du dépôt.

B.2 La base de données MACADAM

B.2.1 Résumé

Les progrès en technique de séquençage et en bioinformatique ont ouvert de nouvelles possibilités, notamment celle de lier les annotations de génome avec de l'information fonctionnelle telles que les données sur les voies métaboliques. Grâce au développement des bases de données d'annotations fonctionnelles, les scientifiques sont capables de lier les annotations des génomes à une information fonctionnelle. Nous présentons ici la base de données MACADAM (pour, en anglais : *MetAboliC pAthways DAtabase for Microbial taxonomic groups*). MACADAM est une base de données qui regroupe les voies métaboliques en associant des statistiques sur leur complétude pour un taxon procaryote donné. Son accès se veut simple et ouvert. Pour chaque « complete genome » procaryote provenant de la base de données RefSeq, MACADAM construit une PGDB (de l'anglais : *Pathway Genome DataBase*) en utilisant le logiciel Pathway Tools qui lui-même se base sur les informations des voies métaboliques contenues dans la base de données MetaCyc. MACADAM récolte également les informations sur les réactions, les enzymes et les métabolites de ces voies métaboliques. Afin de s'assurer que les données fonctionnelles soient de la plus haute qualité possible, MACADAM contient également la base de données MicroCyc, une collection de PGDBs ayant été soumis à un processus de curation manuel ; FAPROTAX (de l'anglais : *Functional Annotation of Prokaryotic Taxa*), une base de données d'annotations fonctionnelles tirées de la littérature sur les organismes cultivables ; et IJSEM phenotypic database, une base de données contenant des données phénotypiques, métaboliques et de tolérance environnemental issues du journal IJSEM de 2004 à 2014. La base de données MACADAM contient 13 509 PGDBs (13 195 PGDBs bactériennes et 314 PGDBs d'archées) et 1 260 voies métaboliques uniques qui sont complétées par 82 annotations fonctionnelles de FAPROTAX et 16 issues de la base phénotypique IJSEM. MACADAM contient un total de 7 921 métabolites, 592 réactions enzymatiques, 2 134 nomenclatures EC et 7 440 enzymes. MACADAM peut être interrogé à n'importe quel rang taxonomique (du phylum à l'espèce) selon la nomenclature NCBI Taxonomy. MACADAM permet d'explorer l'information fonctionnelle de ces taxons complétée par la liste des métabolites, des enzymes, des réactions et des nomenclatures EC. Les résultats de requêtage MACADAM sont produits sous la forme d'un fichier tabulé contenant la liste des voies métaboliques présentes dans le taxon. Chaque voie métabolique est associée à deux scores (PS) : Pathway Score et PFS : Pathway Frequency Score). Le fichier contient également le nom des organismes associés aux

PGDBs utilisées pour décrire le taxon dans lesquels la voie métabolique est identifiée. Chacune des voies est associée à sa classification hiérarchique. MACADAM peut être téléchargée comme un simple fichier, afin d'être librement consultable en local, ou peut être utilisée via une interface web disponible à l'adresse suivante : `http://macadam.toulouse.inra.fr/`. L'ensemble des scripts de construction de MACADAM est disponible à l'adresse suivante : `https://github.com/maloleboulch/MACADAM-Database`.

Le score de complétude des voies métaboliques Chacune des voies métaboliques contenue dans MACADAM est associée avec un score de complétude propre à l'organisme (PS). Ce score est unique à chaque voie métabolique dans un organisme donné. Il représente le nombre de réactions composant une voie métabolique pour un organisme précis. Ce score oscille entre 0 (aucune réaction composant la voie n'est présente dans cet organisme) à 1 (toutes les réactions composant la voie sont présentes). Ce score de complétude ne prend pas en compte les multiples copies d'une même réaction. Cette dernière information est prise en compte dans le score de fréquence de la voie métabolique (PFS). Ce dernier peut aller au delà de 1 car une réaction peut être présente en plusieurs copies dans l'organisme. Une illustration du calcul du PS et du PFS est présentée figure 2 page 107 de l'article de MACADAM.

Les spécificités de MACADAM MACADAM est une base de données de voies métaboliques et d'annotations fonctionnelles sur l'intégralité des phyla procaryotes connus actuellement. A chaque mise à jour, elle intègre les nouvelles versions de RefSeq, du NCBI Taxonomy et de MicroCyc permettant de suivre au plus près les corrections apportées à la taxonomie ainsi que les nouveaux génomes déposés. Les utilisateurs de MACADAM peuvent restreindre leurs recherches à l'aide de plusieurs critères : le nom de la voie recherchée, une catégorie de voie MetaCyc, un PS minimal ou maximal, une enzyme, une nomenclature EC ou encore un nom de réaction. Mais la spécificité la plus importante et unique de MACADAM est l'inférence de voies métaboliques et d'annotations fonctionnelles pour les taxons sans informations fonctionnelles. Quand un nom de taxon connu mais non relié a une information fonctionnelle est soumis, MACADAM interroge alors le taxon de rang supérieur pour déterminer le potentiel fonctionnel des taxons frères du taxon d'intérêt. Le nombre de fois que la voie est retrouvée chez les taxons frère du taxon d'intérêt est indiqué à l'utilisateur dans le fichier de résultats ce qui lui permet de déterminer l'état de conservation de la voie parmi les organismes composant ce taxon supérieur.

DATABASE
The Journal of Biological Databases and Curation

Database, 2019, 1–14
doi: 10.1093/database/baz049
Original article

Original article

The MACADAM database: a MetAboliC pAthways DAtabase for Microbial taxonomic groups for mining potential metabolic capacities of archaeal and bacterial taxonomic groups

Malo Le Boulch[1], Patrice Déhais[2], Sylvie Combes[1] and Géraldine Pascal[1,*]

[1]GenPhySE, Université de Toulouse, INRA, INPT, ENVT, Castanet Tolosan, France and [2]Sigenae Group, GenPhySE, Université de Toulouse, INRA, INPT, ENVT, Castanet Tolosan, France

*Corresponding author: Tel.: +33 (0)5 61 28 51 05; Fax: +33 (0)5 61 28 53 19; Email: geraldine.pascal@inra.fr

Citation details: Le Boulch,M., Déhais,P., Combes,S. *et al.* The MACADAM database: a MetAboliC pAthways DAtabase for Microbial taxonomic groups for mining potential metabolic capacities of archaeal and bacterial taxonomic groups. *Database* (2019) Vol. 2019: article ID baz049; doi:10.1093/database/baz049

Received 12 October 2018; Revised 26 March 2019; Accepted 27 March 2019

Abstract

Progress in genome sequencing and bioinformatics opens up new possibilities, including that of correlating genome annotations with functional information such as metabolic pathways. Thanks to the development of functional annotation databases, scientists are able to link genome annotations with functional annotations. We present MetAboliC pAthways DAtabase for Microbial taxonomic groups (MACADAM) here, a user-friendly database that makes it possible to find presence/absence/completeness statistics for metabolic pathways at a given microbial taxonomic position. For each prokaryotic 'Ref-Seq complete genome', MACADAM builds a pathway genome database (PGDB) using Pathway Tools software based on MetaCyc data that includes metabolic pathways as well as associated metabolites, reactions and enzymes. To ensure the highest quality of the genome functional annotation data, MACADAM also contains MicroCyc, a manually curated collection of PGDBs; Functional Annotation of Prokaryotic Taxa (FAPROTAX), a manually curated functional annotation database; and the IJSEM phenotypic database. The MACADAM database contains 13 509 PGDBs (13 195 bacterial and 314 archaeal), 1260 unique metabolic pathways, completed with 82 functional annotations from FAPROTAX and 16 from the IJSEM phenotypic database. MACADAM contains a total of 7921 metabolites, 592 enzymatic reactions, 2134 EC numbers and 7440 enzymes. MACADAM can be queried at any rank of the NCBI taxonomy (from phyla to species). It provides the possibility to explore functional information completed with metabolites, enzymes, enzymatic reactions and EC numbers. MACADAM returns a tabulated file containing a list of pathways with two scores (pathway score and pathway frequency score) that are present in the queried taxa. The file also contains the names of the organisms in

which the pathways are found and the metabolic hierarchy associated with the pathways. Finally, MACADAM can be downloaded as a single file and queried with SQLite or python command lines or explored through a web interface.

Github URL: https://github.com/maloleboulch/MACADAM-Database

Database URL: macadam.toulouse.inra.fr

Introduction

For many years, *Bergey's Manual of Determinative Bacteriology* (1) and its successor, *Bergey's Manual of Systematic Bacteriology* (2–6), which provides descriptions of the taxonomy, systematics, ecology, physiology and other biological properties of all described prokaryotic taxa, has been the best consensus for an official prokaryotic classification and the best source of information for prokaryotic organisms and taxa. Thanks to advances in genome sequencing and bioinformatics, it is now possible to link genome annotations and functional information. To make this possible, databases have been built and contain metabolic pathways, e.g. series of chemical reactions catalyzed by enzymes within a cell. For instance, the KEGG database (7) can display any of these pathways in a graphical environment. The Human Metabolome DataBase (HMDB) (8) and Reactome (9) are highly curated and complete databases specializing in human metabolism. WikiPathways (10) is an open access collaborative platform containing metabolic pathways across different species. PATRIC (11) is a bacterial database containing >201 000 prokaryotic genomes, each associated with functional information. BioCyc (12) links the genome sequence of an organism to its functional annotation in >14 560 eukaryotic, bacteria and archaea species. All these databases are referred to as pathway genome databases (PGDBs); i.e. they associate the genome sequences with metabolic pathways. Currently, among available databases, some databases are highly curated, including the EcoCyc (13), BsubCyc (14) and HumanCyc (15) databases devoted to *Escherichia coli* K-12, *Bacillus subtilis* or human metabolic pathways, respectively. They are based on the MetaCyc (16) database, which is a highly curated database containing >2666 metabolic pathways throughout the living world. The MicroCyc database (17), based on the MetaCyc database, has been improved by automatic and manual curation by specialized biologists. Finally, some other databases are also curated using functional information from the literature, e.g. FAPROTAX (18) or the IJSEM phenotypic database (19, 20).

Each of these latter databases has limits to link microbial taxonomy to functional information and is not easily downloadable. HMDB, Reactome and WikiPathways are manually and highly curated, but, despite the high quality of their functional information, they cover a small number of organisms [1, 1 and 31 (3 of which are microbes), respectively]. KEGG, despite having >5299 prokaryotic organisms, has moved to a subscription mode and cannot be accessed offline. PATRIC depends on the KEGG pathway data and cannot be downloaded (Table 1). The BioCyc database is a microbial genome web portal that combines thousands of genomes with pathway information, but the BioCyc website uses a subscription model, free access to the derived BioCyc database is limited to a 2-year-old collection of PGDBs and online consultation is limited to a specific number of times per month. Further access requires a paid subscription. MetaCyc contains a greater number of metabolic pathways than KEGG (21) (2666 vs. 530 metabolic pathways, respectively), and it is freely available for academics, but only a few metabolic pathways are retrievable via a taxonomy or an organism name. With existing databases, it is difficult to request up-to-date functional annotations and up-to-date taxonomic lineages for a taxon (e.g. a whole family). But one MACADAM feature is not provided by these databases: the possibility to infer functional information for prokaryotic taxa with no functional information associated with it. That is why we built MACADAM (MetAboliC pAthways DAtabase for Microbial taxonomic groups), a user-friendly database that makes it possible to find presence/absence/-completeness statistics for metabolic pathways at a given archaeal and bacterial taxonomic rank or organism and to be able to infer functional information using the taxonomy for taxa without functional information. MACADAM is not intended to replace existing databases but provides additional information for scientists wishing to better characterize the functional information of all taxonomic groups, from phyla to species.

MACADAM: main characteristics and added value

The following is an introduction to the MACADAM database's main characteristics and improvements compared to existing databases. Its advantages are 4-fold.

Table 1. Overview of MACADAM, BioCyc, PATRIC and KEGG database features with a focus on metabolic pathway and functional information among prokaryotic organisms

	MACADAM	BioCyc	PATRIC	KEGG
Microbial taxonomy used for requests	NCBI taxonomy	NCBI taxonomy	NCBI taxonomy	KEGG taxonomy (with cross-link to NCBI)
Query possibilities	On one or several taxonomies or organisms, with few filters	On one organism, with multiple filters	On one or several taxonomies or organisms, with multiple filters	On one organism, with no filters
Number of bacterial organisms	13 195	~13 400	198 855	5014
Number of archaeal organisms	314	~400	3069	285
Number of unique metabolic pathways	1260	2666	143	530
Genome origins	RefSeq (complete genomes)	Genbank and RefSeq	Genbank and RefSeq	Genbank and RefSeq
Functional annotation sources	RefSeq (functional annotations), MetaCyc (metabolic pathways), MicroCyc (metabolic pathways), FAPROTAX (functional features), IJSME PhenoDB (phenotypic data)	Genbank/RefSeq (annotations) and MetaCyc (metabolic pathways)	Genbank/RefSeq, KEGG (metabolic pathways)	KEGG
Analysis tools and metrics	PS and PFS	SmartTables, genome browser, omics data analysis, metabolic models and routes and comparative analysis	KEGG pathway map, comparative pathway heatmap, multiple sequence alignment, enzymes and genes conservation in pathway	KEGG mapper tools
Output data types	Metabolic pathway name, pathway class hierarchy, hyperlink to MetaCyc pathway and functional information of the upper rank for taxa without data	Metabolic pathway name, pathway class hierarchy, pathway map including enzymes and metabolites, associated genes, protein associated with pathways and literature references	Metabolic pathway name, pathway class hierarchy, KEGG pathway map including enzymes and metabolites, associated genes, enzyme and gene evolution data	Metabolic pathway name, pathway class hierarchy, KEGG pathway map including enzymes and metabolites, associated genes and literature references
Database flat files downloadable	Yes	Yes for academics	No	Yes with license
Results downloadable	Yes	Yes	Yes	Available on the web only
Command line interrogation	SQLite or python script	Application Programming Interface (API)	API + free command line software	API
Frequency of updates	6 months	2–6 months	6 months	3 months

First, the user has the possibility to explore functional traits at different taxonomic levels, from phyla to species, or dedicated strains covering both the *Archaea* and *Bacteria* kingdoms. The user can request data about several taxonomies or organisms at the same time and can do it using partial names. Second, MACADAM relies on up-to-date NCBI taxonomy (22–24). Third, it is based on high-quality genome functional annotations thanks to the aggregation of (i) up-to-date RefSeq genomes annotated with MetaCyc functional data, (ii) high-quality functional annotations from the MicroCyc database (25) and (iii) the FAPROTAX and IJSEM phenotypic databases (manually curated databases). Fourth, MACADAM is open access, freely downloadable and the possibilities of consultation are unlimited. Altogether, MACADAM facilitates taxonomical functional inference studies by synthesizing high-quality and up-to-date functional annotations and taxonomical information in one source. Table 1 shows the comparison between MACADAM, BioCyc, PATRIC and KEGG databases in the context of queries on functional and

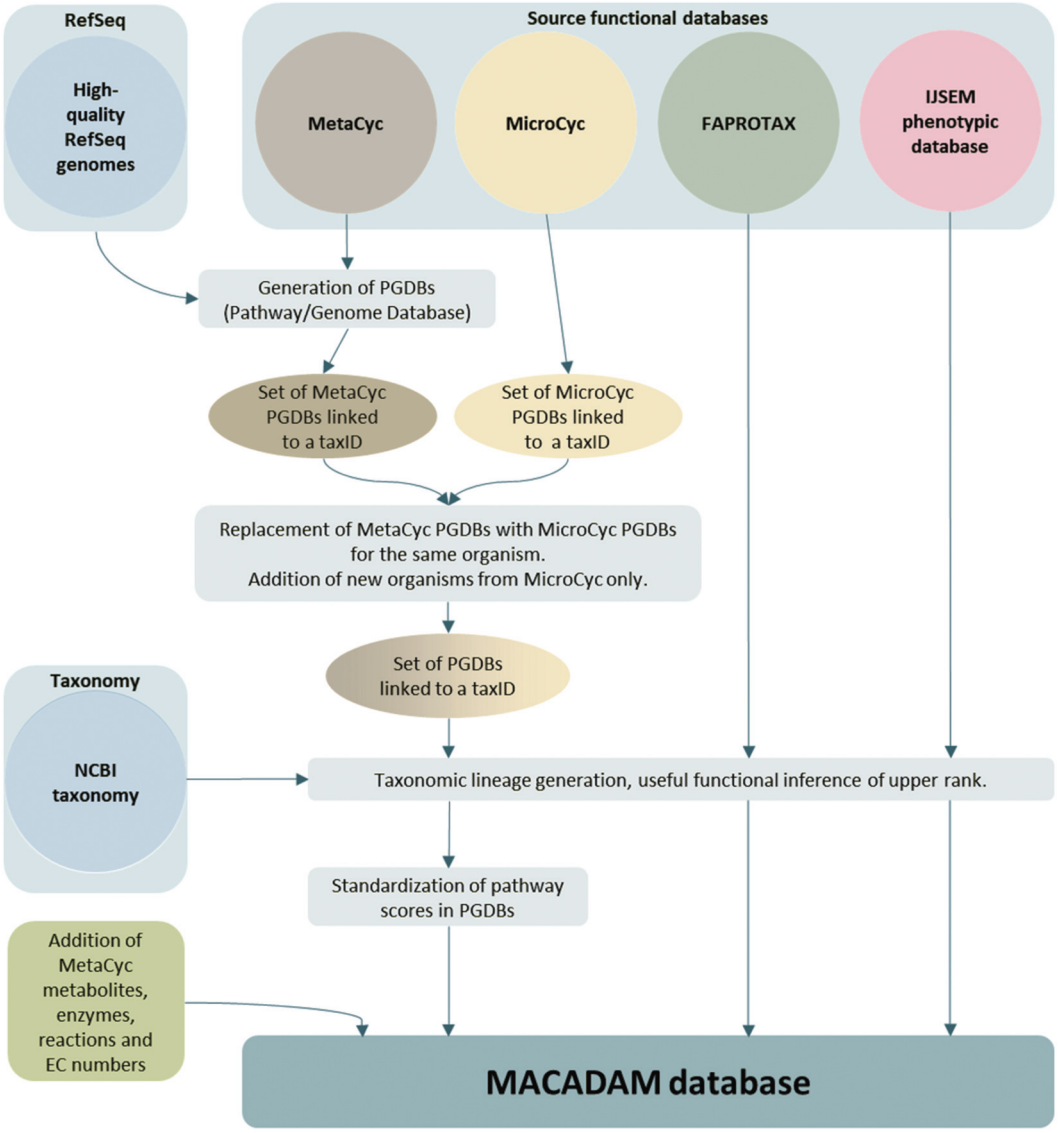


Figure 1. MACADAM building workflow.

metabolic pathway information for microbial organisms. WikiPathways does not appear in Table 1 because it contains only three microbial organisms and it is not useful for microbial queries.

MACADAM content

Building PGDBs between RefSeq and MetaCyc

To build the MACADAM database (Figure 1), we first sourced genomic data of prokaryotes from NCBI RefSeq (25), version 92 (16 January 2019). To ensure high-quality genomic data, we only kept the 11 794 organisms (11 514 bacteria and 280 archaea) with 'complete genome' RefSeq labels (Table 2). These genomes have no gaps, no runs of 10 or more ambiguous bases and the entire chromosome is present (26). This quality allows better annotations using the NCBI prokaryotic genome annotation pipeline (27). Out of the 11 794 genomes, 1683 were labeled 'representative genomes' (1543 bacteria and 140 archaea), i.e. a subset of genomes with an additional quality assurance analysis based on annotation quality metrics. In addition,

Table 2. Statistics on PGDBs collected for MACADAM (values are mean ± SD; values in brackets are minimum and maximum values)

	Number of organisms	Pathways per organism	Number of different pathways	PS	PFS
MetaCyc PGDBs	9954	156 ± 60 [3–350]	851	0.85 ± 0.21 [0–1]	1.37 ± 1.21 [0–51]
MicroCyc PGDBs that have replaced a MetaCyc PGDB	1560	247 ± 80 [26–425]	1012	0.75 ± 0.28 [0.344–1]	1.37 ± 1.45 [0.344–77]
PGDBs only present in MicroCyc	1681	255 ± 70 [2–422]	1007	0.75 ± 0.28 [0.344–1]	1.32 ± 1.28 [0.344–47]
MACADAM bacteria	**13 195**	**179 ± 76 [2–425]**	**1224**	**0.82 ± 0.24 [0–1]**	**1.36 ± 1.26 [0–77]**
MetaCyc PGDBs	184	60 ± 14 [16–91]	207	0.84 ± 0.22 [0.2–1]	1.35 ± 1.11 [0.2–15]
MicroCyc PGDBs that have replaced a MetaCyc PGDB	96	107 ± 25 [2–156]	393	0.77 ± 0.28 [0.05–1]	1.28 ± 1.12 [0.05–21]
PGDBs only present in MicroCyc	34	98 ± 22 [8–149]	344	0.77 ± 0.28 [0.05–1]	1.21 ± 0.93 [0.05–16]
MACADAM archaea	**314**	**79 ± 29 [2–156]**	**478**	**0.80 ± 0.26 [0.05–1]**	**1.31 ± 1.10 [0.05–21]**
MACADAM total	**13 509**	**177 ± 76 [1–425]**	**1260**	**0.82 ± 0.24 [0–1]**	**1.36 ± 1.26 [0–77]**

in bold: summary of metrics for bacteria, archaea or both.

118 bacteria are labeled 'reference genome' (no archaea). They correspond to the highest quality data set, supported by the curation of NCBI scientific staff, and are manually curated. To associate metabolic pathway information with these genomes, we used Pathway Tools software (28) (version 20.5) that relies on MetaCyc data. We thus build a PGDB per organism, i.e. 11 794, that consists of a list of pathways based on the genome annotation. Moreover, to complete pathway information, MACADAM includes the metabolites, reactions, EC numbers, enzymes and hierarchical classification of each pathway. The result is a set of PGDBs and an up-to-date functional annotation of multiple organisms.

Embedding PGDBs from MicroCyc

In the second step of MACADAM building, we added data from MicroCyc (17). MicroCyc is a collection of PGDBs created within the framework of the MicroScope project (29) based on MetaCyc data. These PGDBs are generated from genomes that are (i) manually (re)annoted, (ii) improved thanks to the addition of enzymatic function predictions that are computed with PRIAM software (30) and (iii) completed by annotations from biologist curations using the MaGe system (31). Using the standardized pathway score (PS; Figure 2) proposed in Pathway Tools 16.5, we compare MetaCyc and MicroCyc PGDBs for each organism that is present in both databases. As expected, due to the curation process, MicroCyc PGDBs exhibit more metabolic pathways with a PS equal to 1 than MetaCyc PGDBs (Figure 3). Moreover, MicroCyc contains more pathways per organism than MetaCyc (Table 1). Thus, for common PGDBs between the two databases, those from MicroCyc are chosen over those from MetaCyc and are included in MACADAM, i.e. a total of 1656 PGDBs (1560 bacteria and 96 archaea). MACADAM is then enriched with MicroCyc PGDBs of organisms that are absent from MetaCyc PGDBs, i.e. 1715 (1681 bacteria and 34 archaea). To be added to MACADAM, MicroCyc PGDBs have to be linked to a prokaryotic organism and have a taxonomy recognized in the NCBI taxonomy. As for the MetaCyc process, MACADAM includes the metabolites, reactions, EC numbers, enzymes and hierarchical classification of each pathway. We therefore obtain an improved set of PGDBs from MetaCyc and MicroCyc, completed with functional annotations of 13 509 organisms (Table 1).

Embedding functional annotations from the FAPROTAX and IJSEM phenotypic database

In the third step of MACADAM construction, we added functional traits extracted from the FAPROTAX database (18) and the IJSEM phenotypic database (19). FAPROTAX contains soil and marine bacteria. FAPROTAX maps prokaryotic clades (e.g. genera or species) to establish metabolic or other ecologically relevant functions using the current literature on cultured strains, e.g. *Bergey's Manual of Systematic Bacteriology* (2–6), the Prokaryotes (32) and the IJSEM journal (33). The IJSEM phenotypic database contains phenotypic, metabolic and environmental tolerance data of prokaryotic strains manually extracted from articles in the IJSEM journal. We completed MACADAM with 82 functional annotations from FAPROTAX and 16 from the IJSEM phenotypic database (Table 3). Since this information has a different structure, the PS cannot be calculated.

$$PathwayScore = \mathbf{PS} = \frac{number\ of\ reactions\ that\ have\ an\ annotation\ in\ the\ genome\ for\ a\ pathway}{total\ number\ of\ reactions\ constituting\ the\ pathway}$$

$$PathwayFrequencyScore = \mathbf{PSF} = \frac{number\ of\ enzymes\ annotated\ in\ the\ genome\ for\ a\ pathway}{total\ number\ of\ reactions\ constituting\ the\ pathway}$$

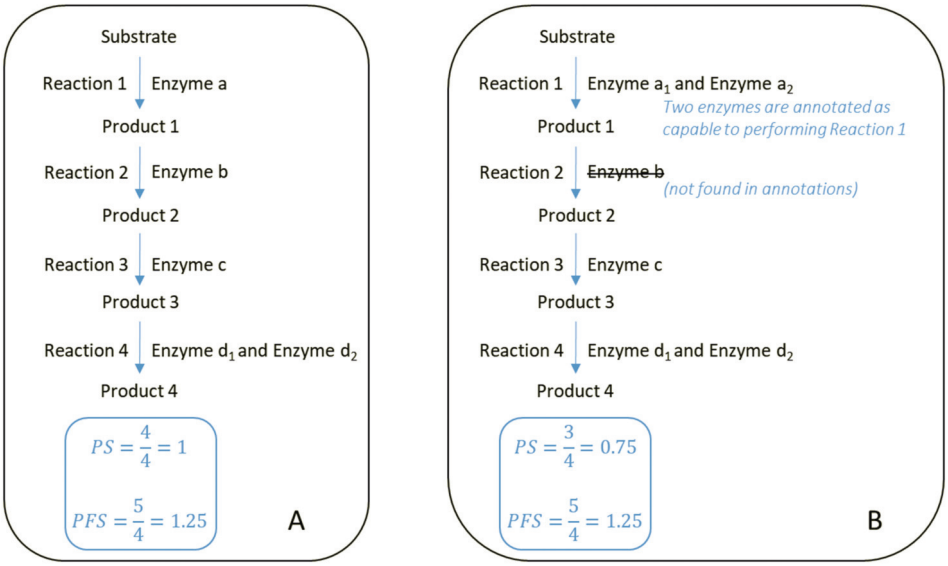


Figure 2. At the top, formulas for the computation of the PS and PFS. Below, examples of two types of computations of the PS and PFS based on an example of a hypothetical metabolic pathway. By comparing the PS and PFS in **A** and **B**, pathway A shows greater evidence of its veracity.

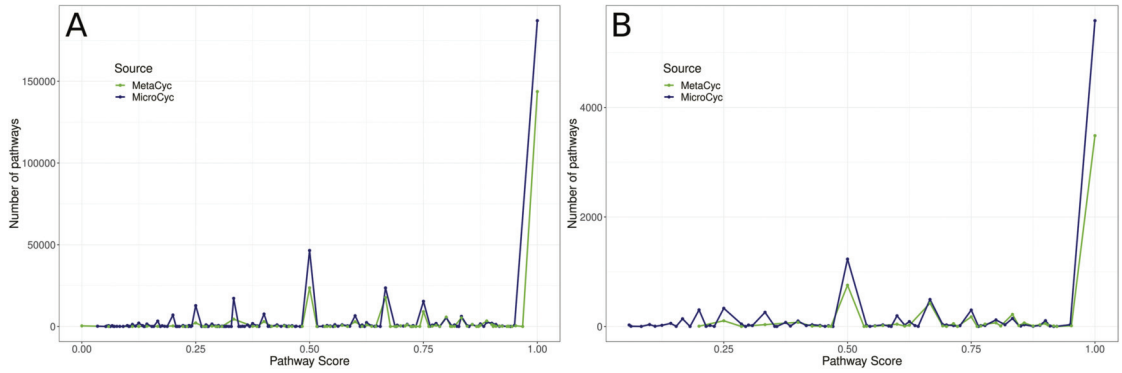


Figure 3. Comparison of PS of all metabolic pathways of PGDBs present in both MetaCyc PGDBs and MicroCyc PGDBs. **A**: among *Bacteria*; **B**: among *Archaea*.

Taxonomy in MACADAM

To link this information, we connected data from PGDBs, the FAPROTAX and the IJSEM phenotypic database with their NCBI taxonomy ID (taxID). This taxID is a unique time-stable identifier because of its interoperability with other taxonomic databases (34). Briefly, in MACADAM, each organism taxonomy has seven taxonomic ranks: superkingdom, phylum, class, order, family, genus and species. Each rank and organism is described by its NCBI taxID. Each MACADAM organism is therefore associated with its numeric lineage formed by seven taxID. For example, *E. coli* is described as *Bacteria*, *Proteobacteria*, *Gammaproteobacteria*, *Enterobacterales*, *Enterobacteriaceae*, and *Escherichia*. *E. coli* is linked to this lineage 2.1224.1236.91347.543.561.562 in MACADAM. However, even if MACADAM data are not associated with minor ranks, e.g. subclasses or subgenera, these minor ranks are kept in MACADAM so that users can find functional information at these minor ranks.

Table 3. Statistics on FAPROTAX and IJSEM phenotypic database information in the MACADAM database for bacterial and archaea organisms

	Number of organisms or taxa	Phenotypic, metabolic or environmental tolerance data
FAPROTAX	*Bacteria*: 3838; *Archaea*: 181	*Bacteria*: 82; *Archaea*: 44
IJSEM phenotypic database	*Bacteria*: 4240; *Archaea*: 87	*Bacteria*: 16; *Archaea*: 12

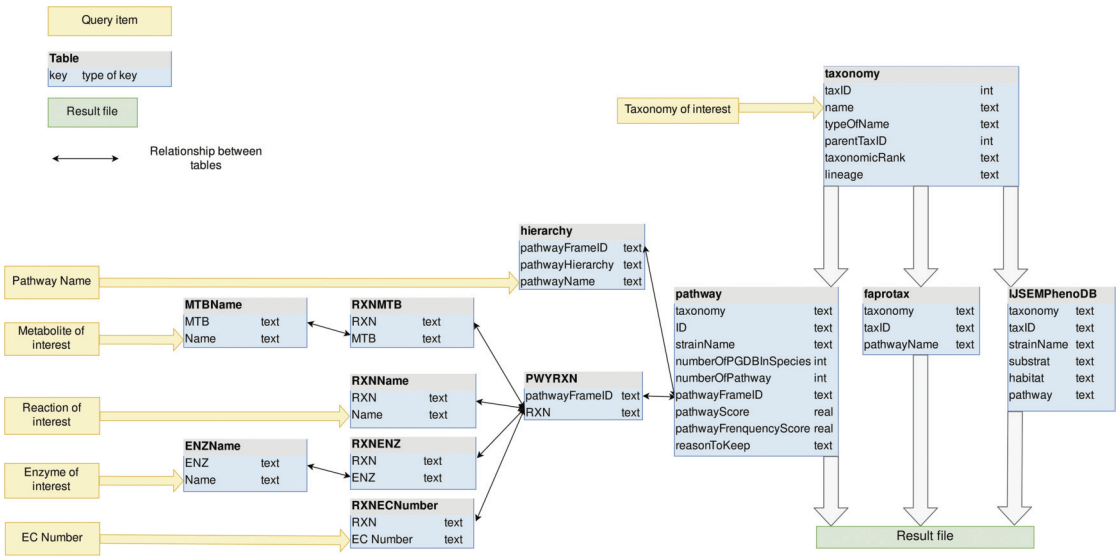


Figure 4. MACADAM database schema. Yellow arrows indicate the entry points of the database.

In conclusion, to create MACADAM, we need (i) RefSeq 'complete genomes', (ii) Pathway Tools with MetaCyc, (iii) MicroCyc PGDBs, (iv) NCBI taxonomy files, (v) the FAPROTAX file and (vi) the IJSEM phenotypic database file.

MACADAM: structure and management

Structure of the MACADAM database

The MACADAM database is a SQLite database (35) that optimizes and facilitates querying. SQLite is an object-relational database management system. It increases the query speed of databases. MACADAM is organized into 12 tables (Figure 4).

As shown in Figure 4, the taxonomy table contains all of the prokaryotic taxonomic names of NCBI taxonomy, whether a PGDB exists or not. Indeed, each PGDB is associated with an organism, i.e. with a species or a strain taxID and PGDBs do not exist at upper taxonomic ranks as genus, family and others. However, to provide users with functional information at these upper ranks, MACADAM needs to store all prokaryotic taxonomies. Thus, the corresponding table in MACADAM contains the taxID of each taxonomic name, the taxID of its parent (parentTaxID) and its taxonomic rank name (taxonomicRank), if any, and the numeric lineage up to this taxID, e.g. taxonomic name = *Proteobacteria*—taxID = 1224, parentTaxID = 2, taxonomicRank = phylum—lineage = 2.1224.

An important table is the pathway table (Figure 4). This table is composed of nine keys. The first one is the taxonomy key that is the numeric lineage. The second one is the ID key that is a unique identifier associated with each PGDB. It consists of the taxID of the corresponding organism, the RefSeq genome label ('rep' for representative genomes, 'ref' for reference genome or 'nan' for complete genomes) and its database of origin (MetaCyc or MicroCyc). The third key is the strainName of a true strain name if it exists. Otherwise, it is provided with the species name. The fourth key is the numberOfPGDBInSpecies. For example, MACADAM has 667 PGDBs linked to *E. coli* species. This means that these 667 PGDBs are linked to 667 strains of *E. coli*; thus, the numberOfPGDBInSpecies for *E. coli* is equal to 667. The fifth key is the number of pathways in a PGDB (numberOfpathway). The sixth key is the pathwayFrameID key that is the official unique

MetaCyc identifier of each pathway. The seventh and eighth keys are PS and the pathway frequency score (PFS), as described in Figure 2. The last key is the ReasonToKeep key that explains why we decided to keep the pathway in the database. The reasons are either that the pathway is from MicroCyc, that the pathway has a high enough threshold quality score, or that the metabolic pathway is complete.

In the hierarchy table, the pathwayHierarchy corresponds to the pathway functional hierarchy found in MetaCyc. For example, for the nitrogen fixation I (pathwayName) pathway, its hierarchy is Degradation/Utilization/Assimilation—Inorganic Nutrient Metabolism—Nitrogen Compound Metabolism—Nitrogen fixation, and its pathwayFrameID key is N2FIX-PWY. The PWYRXN table binds each pathwayFrameID to all MetaCyc official identifiers of reactions that compose the pathways. Each reaction can be described by its name, its metabolites, its enzymes and its EC numbers. These data are stored in RXNName, RXNMTB, RXNENZ and RXNECNumber tables, respectively. For example, the adenosine deoxyribonucleotide *de novo* biosynthesis pathway is composed, among other things, of ADPREDUCT-RXN, known as ADP reductase (EC number = 1.17.4.1), which reduces the ADP metabolite thanks to the ribonucleoside-diphosphate reductase enzyme.

For each organism encoded with its taxID key, the FAPROTAX table and the IJSEMPhenoDB table contain its numeric lineage (taxonomy key) and complementary information such as functional features (pathwayName or pathway keys), environmental habitat or substrate to culture the organism.

Management of the MACADAM database update

How: The MACADAM database is built from a pipeline of python scripts. The update process takes around 2 days, depending on the parallelization capacities. In terms of dependencies, MACADAM requires Python 3, the Pandas package and a valid license of Pathway Tools that have to be installed. All other dependencies are included by default in the python 3 setup. MACADAM can be updated at each RefSeq release (i.e. addition of new high-quality annotated genomes). MACADAM automatically downloads the new index summary from RefSeq. MACADAM then builds script downloads and processes all the genomes that matched our quality standards and launches Pathway Tools on each one of them. This is the crucial part in terms of computing power. In order to save time, the process is parallelized on the cluster of the GenoToul Bioinformatics Platform (http://bioinfo.genotoul.fr/) that provides access to high-performance computing resources. To do this, we process genomes by batches of 50 genomes and the whole step takes ∼1 day. This step needs at least 8 GB of RAM for each batch. The generation of the unique ID and the calculation of the PS take around 6 h each due to file movements and extraction of archives. MACADAM can take up to 600 GB of disk space during the construction of the PGDBs, which is the most critical part. After compression, this file takes up 150 GB of disk space.

Who: MACADAM database benefits from the GenoToul Bioinformatics Platform facilities, including a permanent staff and technical support. The database will be available for downloading and querying as long as it demonstrates its utility for the research community. Moreover, the python script pipeline is available on the GitHub repository (GitHub URL: https://github.com/maloleboulch/MACADAM-Database).

How often: We intend to update MACADAM every 6 months so as to benefit from the latest information on genome annotation from RefSeq and prokaryotic classification from NCBI taxonomy.

Querying the MACADAM database

Input query

MACADAM provides users with metabolic pathways and functional annotations about taxonomic names. Thus, the user has to query MACADAM with one or several species and/or one or several taxonomic names (if several, they have to be separated by a comma). For more precise queries on taxonomic names, the taxonomic rank can be specified thanks to a drop-down menu (phylum, order, class, family, genus or species). Otherwise, the user has to query on 'all ranks'. Thus, if the user specifies 'coli' and 'genus', *E. coli* will not be a match. One specificity of MACADAM is that if the queried organism/taxonomic name has no link to functional information, the user still obtains functional information about the upper rank of the target. To limit this, the 'Strict taxonomy' option allows the user to query only on target organism/taxonomic names and not beyond the specified taxonomic ranks. The search can also be refined by specifying (i) the full name (or part of the name) of a metabolic pathway or a functional feature (e.g. a query with 'nitrate' limits output information to 11 MetaCyc pathway names, 4 functional annotations from FAPROTAX and 1 functional annotation from the IJSEM phenotypic database), (ii) a MetaCyc class hierarchy ID (e.g. a query with 'denitrification' limits the output to 3 MetaCyc metabolic pathways and 4 FAPROTAX functional annotations), (iii) a specific PS (min score = 1 means that MACADAM returns only complete metabolic pathways), (iv) metabolite names (n = 7921), (v) reaction names

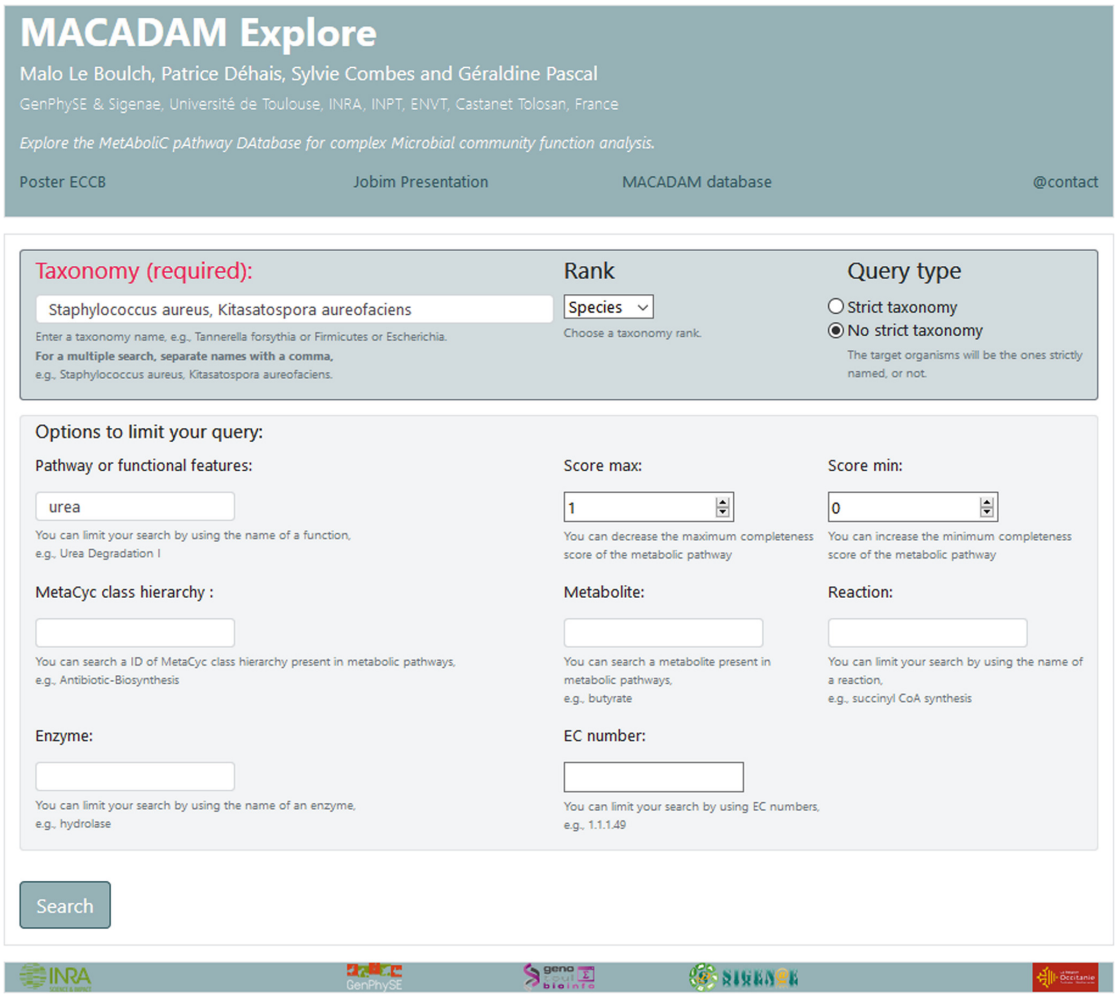


Figure 5. Screenshot of MACADAM Explore website showing the query of all functional information containing the word 'urea' in the species '*Staphylococcus aureus*' and '*Kitasatospora aureofaciens*'.

(n = 592), (vi) enzyme names (n = 7440) and/or (vi) EC numbers (n = 2134). All search fields can be filled with complete or incomplete strings.

Web interface: MACADAM explore

We have created a web interface called MACADAM Explore (Figure 5) to facilitate the consultation of the MACADAM database. MACADAM Explore is built on an Apache Hypertext Transfer Protocol (HTTP) server. It provides commands to search and retrieve the data in the database. Python CGI is used for the front–end web interface. All common gateway interfaces and database interfacing scripts are written in the Python programming language. The database file (SQLite format) with a query script is downloadable at http://macadam.toulouse.inra.fr.

Output file

MACADAM returns a Tabular Separated Value (TSV) file as output (Figure 6) that is downloadable on a personal computer. This output contains a query reminder and details on the matching taxonomy in MACADAM. Moreover, it contains the list of functional information that responds to input criteria. The second column provides the number of organisms that have the targeted pathway. A high proportion shows a high presence of targeted pathways among targeted taxonomic names. For example, as shown in Figure 6A, the 'Urea Degradation II' pathway is

Request: " *Staphylococcus aureus* " and " *Kitasatospora aureofaciens* ", in " species " and with function " urea "

A

Result file for data on organism(s): *Staphylococcus aureus*, *Kitasatospora aureofaciens*, on rank(s):species, with a maximum completeness score of the metabolic pathway of 1.0, with a minimum score of 0.0, and with function(s) containing: "urea"
The following requested taxonomy (TaxID: 1280, Taxonomy: Staphylococcus aureus) is linked to functional information in the MACADAM Database. Number of PGDBs corresponding to this taxID with this filter: 327
Link to the NCBI Taxonomy Database: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1280&lin=s
The following requested taxonomy (TaxID: 1894, Taxonomy: Kitasatospora aureofaciens) is linked to functional information in the MACADAM Database. Number of PGDBs corresponding to this taxID with this filter: 1
Link to the NCBI Taxonomy Database: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1894&lin=s
Functional information from MetaCyc & MicroCyc:

Metabolic Pathway	Present in X org./Total org.	Median score	Median frequency	Metabolite	Reaction	Enzyme	EC number	Target taxonomies	Strain with the Pathway	Pathway Hierarchy	MetaCyc URL
allantoin degradation to ureidoglycolate I (urea producing)	1/328	1	1	NA	NA	NA	NA	Kitasatospora aureofaciens	Kitasatospora aureofaciens DM-1	Degradation.AMINE-DEG.Allantoin-degradation.PWY-5697.	https://metacyc.org/META/new-image?type=PATHWAY&object=PWY-5697
urea degradation I	9/328	0.5	1.5	NA	NA	NA	NA	Staphylococcus aureus	Staphylococcus aureus subsp. aureus TW20, Staphylococcus aureus CA-347, ...[1]	Degradation.AMINE-DEG.Urea-Degradation.PWY-5703.	https://metacyc.org/META/new-image?type=PATHWAY&object=PWY-5703
urea degradation II	328/328	1	3	NA	NA	NA	NA	Staphylococcus aureus, Kitasatospora aureofaciens	Staphylococcus aureus V521, Staphylococcus aureus ST20130942, ...[1]	Degradation.AMINE-DEG.Urea-Degradation.PWY-5704.	https://metacyc.org/META/new-image?type=PATHWAY&object=PWY-5704
urea cycle	313/328	0.8	1	NA	NA	NA	NA	Staphylococcus aureus, Kitasatospora aureofaciens	Staphylococcus aureus V521, Staphylococcus aureus ST20130942, ...[1]	Degradation.Noncarbon-Nutrients.NITROGEN-DEG.PWY-4984.	https://metacyc.org/META/new-image?type=PATHWAY&object=PWY-4984

Request: "Lactobacillus cerevisiae", in " species " and with function " urea "

B

##Result file for data on organism(s):Lactobacillus cerevisiae, on rank(s):species, with a maximum completeness score of the metabolic pathway of 1.0, with a minimum score of 0.0 and with function(s) containing: "urea"
The requested taxonomy (TaxID: 1704076, Taxonomy: Lactobacillus cerevisiae) is not linked to functional information in the MACADAM Database but the functional information for a upper rank of its lineage is displayed: 1578, Taxonomy: Lactobacillus. Number of PGDBs corresponding to this taxID with this filter: 187
##Link to the NCBI Taxonomy Database: https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1578&lin=s
Functional information from MetaCyc & MicroCyc:

Metabolic Pathway	Present in X org./Total org.	Median score	Median frequency	Metabolite	Reaction	Enzyme	EC number	Target taxonomies	Strain with the Pathway	Pathway Hierarchy	MetaCyc URL
allantoin degradation to ureidoglycolate I (urea producing)	1/187	0.5	0.5	NA	NA	NA	NA	Lactobacillus cerevisiae	Lactobacillus reuteri TD1, Lactobacillus reuteri mlc3, ...[1]	Degradation.AMINE-DEG.Allantoin-degradation.PWY-5697.	https://metacyc.org/META/new-image?type=PATHWAY&object=PWY-5704
urea degradation I	5/187	0.5	1	NA	NA	NA	NA	Lactobacillus cerevisiae	Lactobacillus brevis subsp. gravesensis ATCC 27305, ...[1]	Degradation.AMINE-DEG.Urea-Degradation.PWY-5703.	https://metacyc.org/META/new-image?type=PATHWAY&object=PWY-5697
urea degradation II	10/187	1	3	NA	NA	NA	NA	Lactobacillus cerevisiae	Lactobacillus plantarum LZ95, Lactobacillus backii TMW 1.1988, ...[1]	Degradation.AMINE-DEG.Urea-Degradation.PWY-5704.	https://metacyc.org/META/new-image?type=PATHWAY&object=PWY-4984
urea cycle	180/187	0.6	1.2	NA	NA	NA	NA	Lactobacillus cerevisiae	Lactobacillus fermentum F-6, Lactobacillus mucosae LM1, ...[1]	Degradation.Noncarbon-Nutrients.NITROGEN-DEG.PWY-4984.	https://metacyc.org/META/new-image?type=PATHWAY&object=PWY-5703

Figure 6. Examples of output files corresponding to requests on MACADAM. (**A**) The user has searched for all metabolic pathways in *S. aureus* and *K. aureofaciens*, using the term 'urea' in the function field text. (**B**) The user has searched for all metabolic pathways in *Lactobacillus cerevisiae* using the term 'urea' in the function field text. Since there is no data on this organism in MACADAM, the information was searched for higher up in the taxonomy hierarchy, i.e. *Lactobacillus*. [1]List of organisms in MACADAM with the targeted metabolic pathway.

present in all 328 genomes that match taxID 1280 and 1894 corresponding to *Staphylococcus aureus* and *Kitasatospora aureofaciens*, although the 'Urea Degradation I' pathway is present in only nine genomes of *S. aureus*, and the 'allantoin degradation to ureidoglycolate' is only present in one strain of *K. aureofaciens*. The output also provides information about the median value of the PS and PFS. There are also columns corresponding to metabolite names, reaction names, enzyme names and EC numbers entered as filtered input by the user. The last columns contain targeted taxonomy names, the list of strains that have the pathways and the MetaCyc metabolic pathway hierarchy with the corresponding URL.

Utility and discussion

MACADAM is designed to collect bacterial and archaeal genomes of the highest quality associated with the highest quality annotations Reliable data are needed to infer the functional potential of complex prokaryotic communities. Indeed, obsolete functional information can lead to inaccurate insights (36). As far as possible, MACADAM avoids sequentially spurious annotations. In the MACADAM database, reliability is ensured by filters on genome quality, meaning that only complete genomes have been taken from RefSeq. Moreover, to ensure up-to-date annotations, we compute PGDBs with Pathway Tools at each release, based on RefSeq, NCBI taxonomy, MetaCyc, MicroCyc, FAPROTAX and the IJSEM phenotypic database.

MACADAM PGDBs cover all phyla recognized by the List of Prokaryotic names with Standing in Nomenclature (LPSN; http://www.bacterio.net/) and the 10 other newly proposed phyla from the NCBI taxonomy (Figure 7). Proteobacteria is the most prevalent phylum and accounts for >55% of the genomes collected in MACADAM, followed by the *Firmicutes* and *Actinobacteria* phyla that account for >22% and 11% of the collected genomes, respectively. This pre-eminence is probably explained by the research effort devoted to these phyla by biologists. Accordingly, *Escherichia* and *Salmonella* are the prevalent genera in the database (5.1% and 4.3% of the database, respectively). Interestingly,

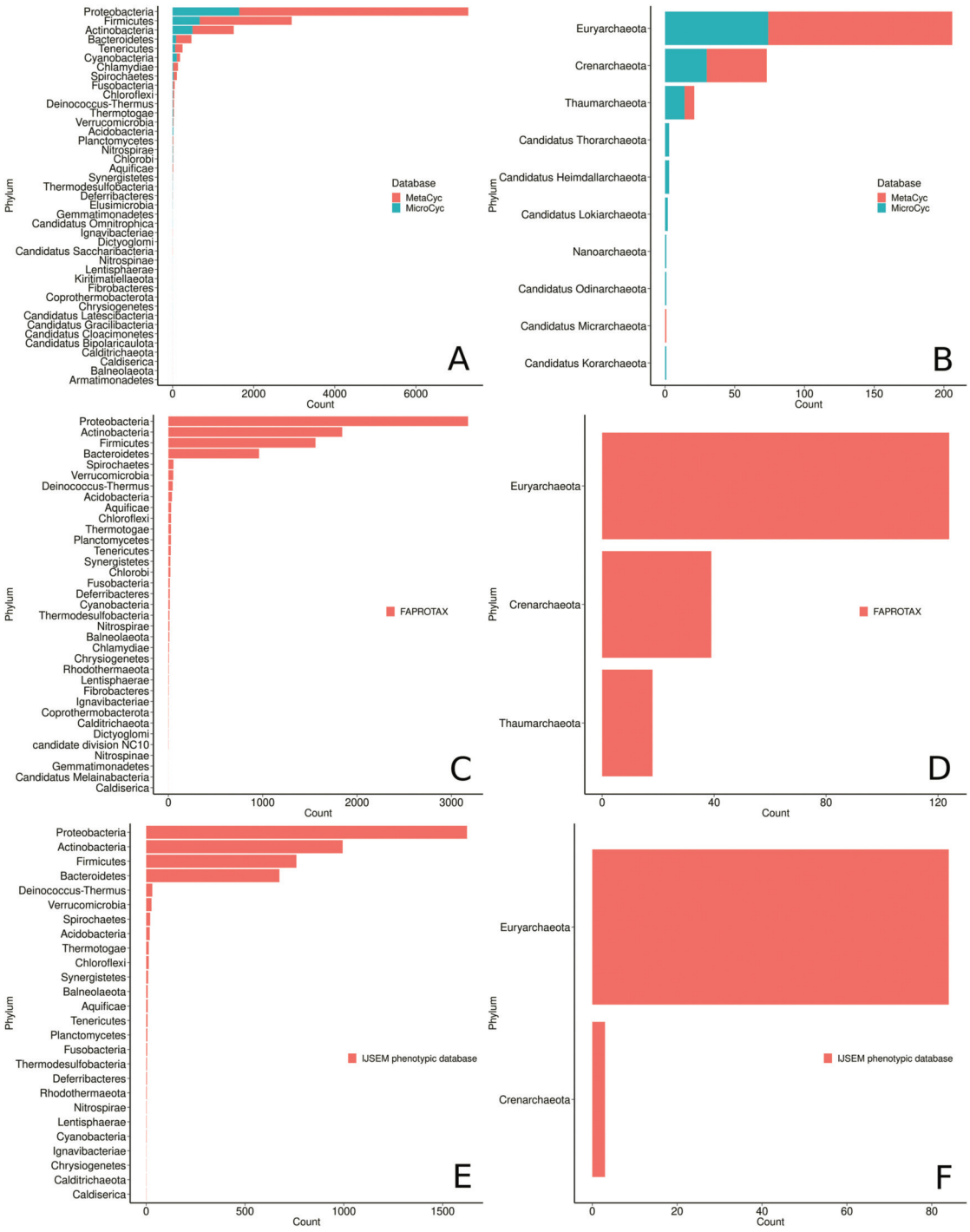


Figure 7. Phyla distribution in the MACADAM database according to their database of origin. (**A**) MetaCyc and MicroCyc for bacterial organisms, (**B**) MetaCyc and MicroCyc for archaea organisms, (**C**) FAPROTAX for bacterial organisms, (**D**) FAPROTAX for archaea organisms, (**E**) IJSEM phenotypic database for bacterial organisms and (**F**) IJSEM phenotypic database for archaea organisms.

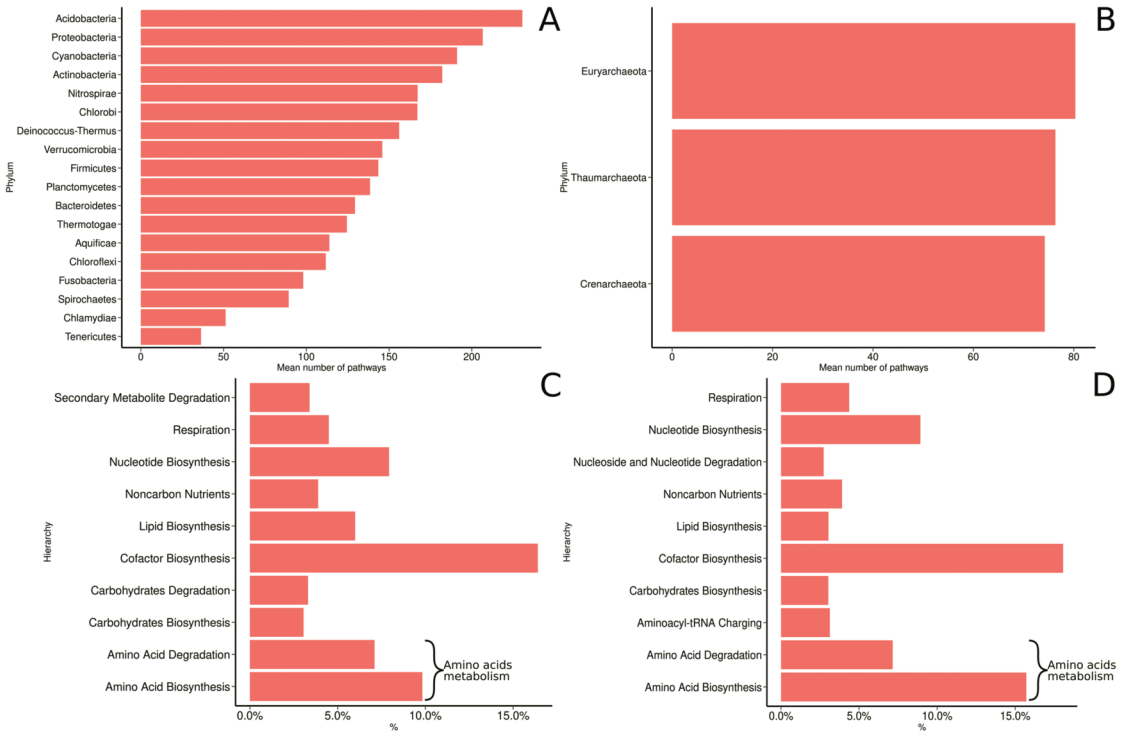


Figure 8. MACADAM functional diversity for phyla with >10 PGDBs in MACADA (**A**) among bacterial phyla, (**B**) among archaea phyla, (**C**) the 10 main hierarchical groups of pathways in all bacterial organisms and (**D**) the 10 main hierarchical groups of pathways in all archaea organisms.

according to Figure 7, the *Acidobacteria* phylum is weakly represented in MACADAM (only 21 PGDBs, 19 of which are from MicroCyc). But these PGDBs include the highest number of metabolic pathways (mean = 230; min = 138; max = 298; Figure 8). Acidobacteria is one of the most widespread phyla, but few organisms of this phylum are cultivated and sequenced (37) and may explain the low representation of available high-quality genome that could be included in MACADAM. The high number of metabolic pathway present in these widespread phyla may indicate that, despite few sequenced organisms, these are subject to particular care in terms of functional annotation. In parallel, we should also recall that the most common metabolic pathways in MACADAM belong to cofactor biosynthesis (bacteria: 16.4%; archaea: 18.06%) and the metabolism of amino acids (bacteria: 17.7%; archaea: 22.86%).

MACADAM allows users to refine queries using PSs PSs range from 0 to 1 (Table 2). A value of 1 indicates that all the enzymes required for the pathway are present in the genome of the organism. A value of 0 indicates that none of the enzymes are present in the targeted genome. In the latter case, the pathway is still present because of the functional inference applied by MetaCyc experts, information that they cross-reference with phenotypic evidence described in the literature. MACADAM contains eight pathways in 1519 organisms with a PS equal to 0. The PFS ranges from 0 to 77 (Table 2). *Nocardia nova SH22a* has the maximum PFS for its long-chain fatty acid activation pathway. This pathway comprises only one reaction and is part of a longer lipid biosynthesis pathway.

Table 4 is an example of statistics performed on L-lysine fermentation to acetate and butanoate pathway output (MetaCyc ID: P163-PWY), a pathway of interest in understanding the interactions between host and gut microbiota in health and disease. In fact, butyrate is a microbial fermentation product that is used as an energy source by enterocytes and whose signaling properties are involved in multiple functions of enterocytes, including cell differentiation, gut tissue development, immune modulation, oxidative stress reduction and diarrhea control (38). Ten enzymes are involved in the L-lysine fermentation to acetate and butanoate pathway. The pathway is present in 992 different organisms in MACADAM. This complete pathway is composed of 10 enzymes. If the median value of the PS is equal to 0.4, this means that four reactions have at least one annotated enzyme in the pathway. If the median value of

Table 4. Statistics on the L-lysine fermentation to acetate and butanoate pathway (MetaCyc ID is P163-PWY; values in brackets are minimum and maximum values

MetaCyc ID: P163-PWY characteristics	
Number of reactions in the complete pathway*	10
Number of bacteria in which this pathway is present	756
Median number of unique enzymes present in this pathway in organisms	4 [1–10]
Median number of enzymes present in this pathway in organisms	7 [1–70]
Key reaction*	1 (enzyme classification: 5.4.3.2)
PS	0.4 [0.1–1]
PFS	0.7 [0.1–7]

*Value found in a MetaCyc flat file; a key reaction is a reaction that is specific to a single pathway, i.e. a reaction that is not found in any other pathway.

the PFS is equal to 0.7, this means these 4 reactions have >1 associated enzymes, 7 enzymes for 10 reactions in all (Figure 2). According to MetaCyc, the 5.4.3.2 reaction is a key reaction in this pathway. Therefore, if the organism contains this reaction, the whole pathway will be identified for the organism. These PS and PFS data are useful to biologists for data mining.

An important feature of MACADAM is its ability to **infer functional annotations for taxa with no associated genomic sequences using the taxonomy**. In the case of a taxon with no functional information in MACADAM, i.e. missing in MetaCyc, MicroCyc, FAPROTAX and the IJSEM phenotypic database, we provide information for the upper taxonomy rank. For example, if a species has no functional information, MACADAM automatically requests functional information at the genus level, just like FAPROTAX. However, MACADAM can do this at any taxonomic rank, while FAPROTAX is limited to the order rank. In MACADAM, in this case, all annotations for organisms belonging to this genus are shown in the output file. As for FAPROTAX, it indicates functional information described in the literature at the genus rank if all described species of the genus have been shown to exhibit the given functional information and not the addition of functional information about all of the organisms belonging to this rank. Thanks to the column that gives the number of organisms with the targeted pathway (column 2 in Figure 6B), it is possible to see pathways that are more or less conserved in the taxon of interest. For example, 'urea cycle' is a pathway conserved in most of *Lactobacillus*, unlike the 'urea degradation I or II' pathways. Thus, this feature provides functional information about organisms with no functional information based on related taxonomic species.

Conclusions

MACADAM was designed for the microbiology community as a functional annotation information database based on multiple sources of data on functional annotations and on metabolic pathways (MetaCyc, MicroCyc, FAPROTAX and the IJSEM phenotypic database). The database is also based on the complete and interoperable NCBI taxonomy. MACADAM covers all known bacterial and archaeal phyla, as of February 2019. A standardized score enables quick comparison and comprehension of the potential presence of a pathway. If there is no functional information on the taxonomy entered, MACADAM automatically checks the upper taxonomic rank in order to provide functional information associated with related organisms to users. MACADAM can be explored via metabolites, reactions, enzymes, EC numbers or specific pathways. A user-friendly web interface makes querying easy. MACADAM will be useful to all biologists who need to determine the functional potential of a prokaryotic species or any other taxonomic rank. Since the source code to build MACADAM is available to everyone (GitHub URL: https://github.com/maloleboulch/MACADAM-Database), MACADAM can be included in any functional inference tool able to integrate the abundance tables of complete microbial communities generated, among others, we plan to include MACADAM in the FROGS software (39) to analyze amplicon metagenomics data.

Acknowledgements

The authors are grateful to the GenoToul Bioinformatics Platform, Toulouse, Occitanie, for providing assistance, computing and storage resources. The authors thank Cédric Cabau and Philippe Bardou for their help with the website. The authors are grateful to Gail Wagman for her correction of the English-language version.

Funding

French National Institute for Agricultural Research (PHASE division); Region Occitanie (R9090198/00100870 to M.L.B.).

Conflict of interest. None decared.

References

1. Bergey,D.H., Buchanan,R.E., Gibbons,N.E. *et al.* (1974) *Bergey's Manual of Determinative Bacteriology*. Williams & Wilkins Co, Baltimore.

2. Garrity,G., Boone,D.R. and Castenholz,R.W. (eds) (2001) *Bergey's Manual of Systematic Bacteriology: The Archaea and the Deeply Branching and Phototrophic Bacteria*, Vol. **1**, 2nd edn. Springer-Verlag, New York.
3. Garrity,G. (ed) (2005) *Bergey's Manual of Systematic Bacteriology: The Proteobacteria*, Vol. **2**, 2nd edn. Springer, USA.
4. Vos,P., Garrity,G., Jones,D. *et al.* (eds) (2009) *Bergey's Manual of Systematic Bacteriology: The Firmicutes*, Vol. **3**, 2nd edn. Springer-Verlag, New York.
5. Krieg,N.R., Ludwig,W., Whitman,W. *et al.* (eds) (2010) *Bergey's Manual of Systematic Bacteriology: The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes*, Vol. **4**, 2nd edn. Springer-Verlag, New York.
6. Whitman,W., Goodfellow,M., Kämpfer,P. *et al.* (eds) (2012) *Bergey's Manual of Systematic Bacteriology: The Actinobacteria*, Vol. **5**, 2nd edn. Springer-Verlag, New York.
7. Kanehisa,M., Furumichi,M., Tanabe,M. *et al.* (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
8. Wishart,D.S., Feunang,Y.D., Marcu,A. *et al.* (2018) HMDB 4.0: the Human Metabolome Database for 2018. *Nucleic Acids Res.*, **46**, D608–D617.
9. Fabregat,A., Jupe,S., Matthews,L. *et al.* (2018) The Reactome Pathway Knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.
10. Slenter,D.N., Kutmon,M., Hanspers,K. *et al.* (2018) WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.*, **46**, D661–D667.
11. Wattam,A.R., Davis,J.J., Assaf,R. *et al.* (2017) Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res.*, **45**, D535–D542.
12. Karp,P.D., Billington,R., Caspi,R. *et al.* (2017) The BioCyc collection of microbial genomes and metabolic pathways. *Brief. Bioinform.* doi:10.1093/bib/bbx085.
13. Karp,P.D., Weaver,D., Paley,S. *et al.* (2014) The EcoCyc database. *EcoSal Plus*, **6**. doi:10.1128/ecosalplus.ESP-0009-2013.
14. Caspi,R., Altman,T., Billington,R. *et al.* (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **42**, D459–D471.
15. Romero,P., Wagg,J., Green,M.L. *et al.* (2004) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.*, **6**, R2.
16. Caspi,R., Billington,R., Fulcher,C.A. *et al.* (2018) The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **46**, D633–D639.
17. Vallenet,D., Calteau,A., Cruveiller,S. *et al.* (2017) MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes. *Nucleic Acids Res.*, **45**, D517–D528.
18. Louca,S., Parfrey,L.W. and Doebeli,M. (2016) Decoupling function and taxonomy in the global ocean microbiome. *Science*, **353**, 1272–1277.
19. Barberán,A., Velazquez,H.C., Jones,S. *et al.* (2017) Hiding in plain sight: mining bacterial species records for phenotypic trait information. *mSphere*, **2**, e00237–e00217.
20. Barberan,A. (2016) International Journal of Systematic and Evolutionary Microbiology (IJSEM) phenotypic database.
21. Altman,T., Travers,M., Kothari,A. *et al.* (2013) A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics*, **14**, 112.
22. Federhen,S. (2012) The NCBI taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
23. Sayers,E.W., Barrett,T., Benson,D.A. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
24. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J. *et al.* (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.
25. Tatusova,T., Ciufo,S., Fedorov,B. *et al.* (2014) RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.*, **42**, D553–D559.
26. Pruitt,K.D., Tatusova,T., Brown,G.R. *et al.* (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
27. Tatusova,T., DiCuccio,M., Badretdin,A. *et al.* (2016) NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.*, **44**, 6614–6624.
28. Karp,P.D., Latendresse,M., Paley,S.M. *et al.* (2016) Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. *Brief. Bioinform.*, **17**, 877–890.
29. Vallenet,D., Engelen,S., Mornico,D. *et al.* (2009) MicroScope: a platform for microbial genome annotation and comparative genomics. *Database (Oxford)*, **2009**. https://doi.org/10.1093/database/bap021.
30. Claudel-Renard,C., Chevalet,C., Faraut,T. *et al.* (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.*, **31**, 6633–6639.
31. Vallenet,D., Labarre,L., Rouy,Z. *et al.* (2006) MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res.*, **34**, 53–65.
32. Rosenberg,E., DeLong,E.F., Stackebrandt,E. *et al.* (eds) (2013) *The Prokaryotes: Prokaryotic Biology and Symbiotic Associations*, 4th edn. Springer-Verlag, Berlin Heidelberg.
33. *Int. J. Syst. Evol. Microbiol.*, https://ijs.microbiologyresearch.org/content/journal/ijsem.
34. Balvočiūtė,M. and Huson,D.H. (2017) SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare? *BMC Genomics*, **18**, 114.
35. *SQLite Home Page* https://www.sqlite.org/index.html (5 February 2019, date last accessed).
36. Wadi,L., Meyer,M., Weiser,J. *et al.* (2016) Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods*, **13**, 705–706.
37. Kielak,A.M., Barreto,C.C., Kowalchuk,G.A. *et al.* (2016) The ecology of acidobacteria: moving beyond genes and genomes. *Front. Microbiol.*, **7**, 744.
38. Bui,T.P.N., Ritari,J., Boeren,S. *et al.* (2015) Production of butyrate from lysine and the Amadori product fructoselysine by a human gut commensal. *Nat. Commun.*, **6**, 10062.
39. Escudié,F., Auer,L., Bernard,M. *et al.* (2018) FROGS: Find, Rapidly, OTUs with Galaxy Solution. *Bioinformatics*, **34**, 1287–1294.

B.3 MACADAMExplore et son utilisation

B.3.1 Introduction

MACADAM est une base de données construite et interrogeable en langage SQLite. Ce mode de requêtage est rapide mais assez complexe pour les non-initiés. Les fichiers constituant la base de données sont téléchargeables via le site web `http://macadam.toulouse.inra.fr/`. Afin de faciliter l'interrogation de MACADAM, ce site web propose un formulaire de requête qui appelle un script d'interrogation nommé MACADAMExplore. Ce script peut être utilisé en parfaite indépendance de l'interface web. Il est utilisable en ligne de commande en utilisant Python. Un fichier de documentation lui est également joint, afin de renseigner l'intégralité des options disponibles lors de l'interrogation de MACADAM. Ce fichier explique également l'interrogation en mode graphique de MACADAM via l'utilisation du logiciel *DB Browser for SQLite*. L'ensemble des scripts ainsi que la documentation sont également disponibles sur `https://github.com/maloleboulch/MACADAMExplore`.

Dans ce chapitre nous présentons deux exemples d'application de MACADAMExplore réalisés dans le cadre de ce travail de thèse : (i) le premier ayant permis la production de données d'inférence fonctionnelle dans le cadre d'une étude dont les résultats ont été publiés dans le Journal Fontiers in Microbiolgy : Read (...) Le Boulch, (...), « Diversity and Co-occurrence Pattern Analysis of Cecal Microbiota Establishment at the Onset of Solid Feeding in Young Rabbits », 2019. (ii) Le second exemple d'application ayant permis de compléter l'étude taxonomique réalisée sur les espèces métagénomiques (Partie B.1, page 79) en étudiant le potentiel fonctionnel des quatre genres concernés par cette étude : *Blautia*, *Mediterraneibacter*, *Megasphaera* et *Ruminococcus*.

B.3.2 Inférence fonctionnelle des groupes taxonomiques bactériens dominants du microbiote cæcal de jeunes lapereaux

MACADAM a été conçu pour explorer le potentiel fonctionnel des taxons procaryotes, pour les rangs taxonomiques allant de l'espèce au phylum. Dans le cadre de l'étude portant sur l'exploration de la diversité et l'identification de co-occurrences d'espèces bactériennes dans le microbiote cæcal chez les jeunes lapins (Read et al. 2019), nous avons comparé le potentiel fonctionnel des quatre groupes taxonomiques dominants.

B.3.2.1 Matériels et méthodes

MACADAM a été exploré dans le cadre de cette étude afin de déterminer le potentiel fonctionnel de quatre taxons majeurs identifiés dans le microbiote cæcal de lapereaux : trois familles majeures (*Lachnospiraceae*, *Ruminococcaceae* et *Eubacteriaceae*) ainsi qu'un genre dominant (*Bacteroides*). Les voies métaboliques présentes dans une majorité d'organismes et pouvant jouer un rôle dans la dégradation des nutriments ont été analysées. Afin d'extraire l'information fonctionnelle des 4 taxons, le script MACADAMExplore a été utilisé, suivi de scripts python afin de rassembler l'information et de calculer la moyenne et l'écart-type des scores PS. Les graphiques ont été produits en utilisant R.

B.3.2.2 Résultats et discussion

Les analyses ont été effectuées en utilisant la version de MACADAM disponible en novembre 2018. Le genre *Bacteroides*, et les familles *Lachnospiraceae*, *Ruminococcaceae* et *Eubacteriaceae* comprenaient 22, 24, 28 et 7 organismes respectivement (Tableau B.9). La figure B.X présente les résultats obtenus dans ces taxons pour les voies appartenant au métabolisme de la dégradation et de la fermentation des polymères, ainsi que la dégradation des glucides. Quatre voies de dégradation des glucides (fucose, xylose, D-mannose et melibiose degradation) ont été identifiées dans 90 à 100 % des 22 organismes appartenant au genre *Bacteroides*. Les bactéries du genre *Bacteroides*, majoritaires chez le lapereau en allaitement exclusif, sont ainsi capables de dégrader les glucides du lait (par exemple le fucose) mais présentent également les voies métaboliques nécessaires à la dégradation des glucides d'origine végétale. Il est intéressant de noter que par rapport aux trois autres groupes dominants de l'écosystème, *Bacteroides* est le seul à pouvoir atteindre l'ensemble de la voie de dégradation II du N-acétylneuraminate et de la N-acétylmannosamine. N-acetylneuraminic est une des formes de l'acide sialicilique, une vaste famille de composés à neuf carbones, présents à la surface des cellules de certaines bactéries et qui jouent un rôle important dans leur habilité à coloniser l'écosystème digestif des mammifères et à y perdurer (Severi, Hood et Thomas 2007).

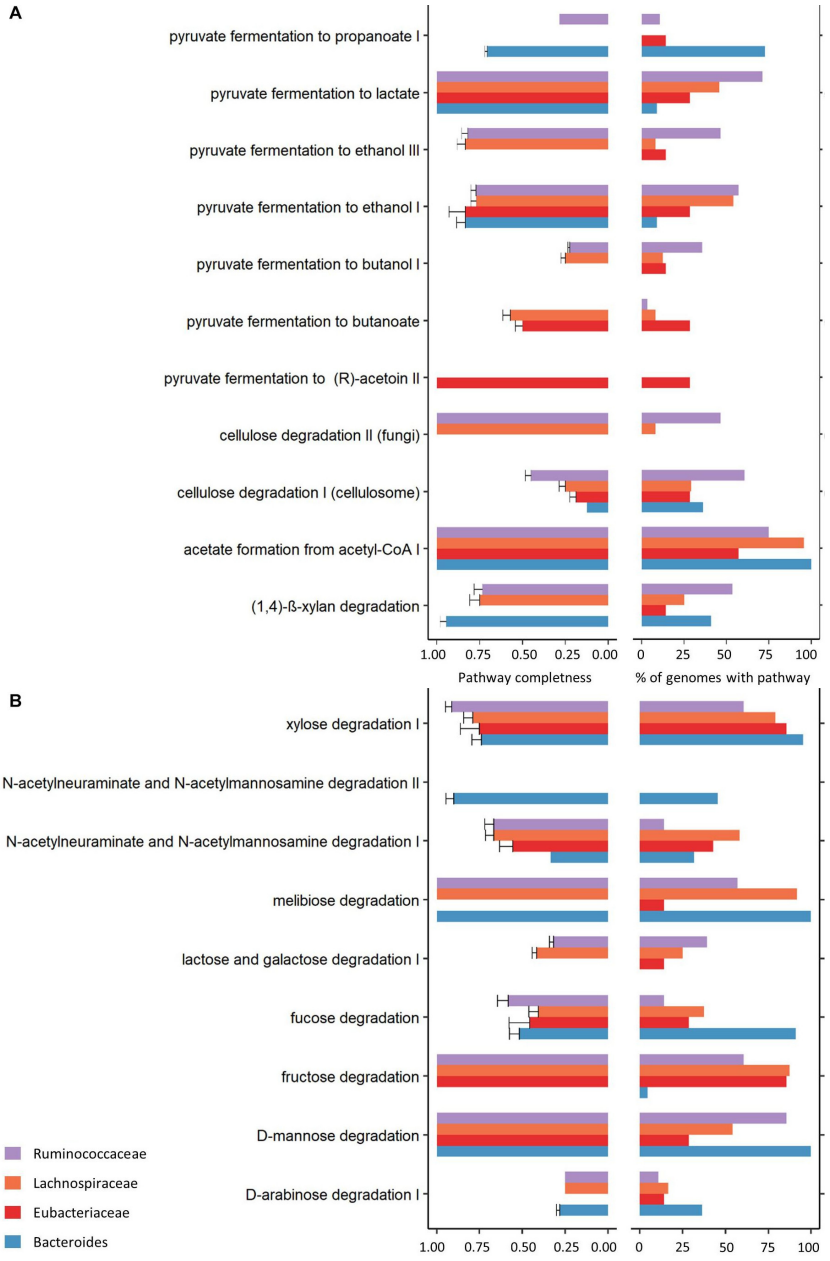


FIGURE B.X – Complétude et pourcentage d'organismes où les voies métaboliques de dégradation ou fermentation de polymères (**A**) ou de dégradation des glucides (**B**) sont présentes, chez les *Lachnospiraceae*, *Ruminococcaceae*, *Eubacteriaceae* et *Bacteroides*. Un score de 1 indique que l'ensemble des réactions composant la voie métabolique est présent dans l'organisme. Un score de zéro indique qu'aucune réaction n'est présente.

Source: READ et al. (2019)

Bacteroides	*Lachnospiraceae*	*Ruminococcaceae*
Bacteroides caccae ATCC 43185	*[Clostridium] bolteae* ATCC BAA-613	*[Clostridium] cellulolyticum* H10
Bacteroides caecimuris I48	*[Clostridium] saccharolyticum* K10	*[Clostridium] cellulosi*
Bacteroides cellulosilyticus WH2	*[Clostridium] saccharolyticum* WM1	*[Clostridium] clariflavum* DSM 19732
Bacteroides dorei CL03T12C01	*[Eubacterium] rectale* ATCC 33656	*[Clostridium] stercorarium* subsp. leptospartum DSM 9219
Bacteroides fragilis 638R	*Anaerostipes hadrus* BPB5	*[Clostridium] stercorarium* subsp. stercorarium DSM 8532
Bacteroides fragilis BE1	*Anaerotignum propionicum* DSM 1682	*[Clostridium] stercorarium* subsp. thermolacticum DSM 2910
Bacteroides fragilis BOB25	*Blautia hansenii* DSM 20583	*[Eubacterium] siraeum* 70/3
Bacteroides fragilis NCTC 9343	*Blautia* sp. YL58	*Acetivibrio cellulolyticus* CD2
Bacteroides fragilis S14	*Butyrivibrio hungatei* MB2003	*Clostridium leptum* DSM 753
Bacteroides fragilis YCH46	*Butyrivibrio proteoclasticus* B316	*Clostridium papyrosolvens* C7
Bacteroides helcogenes P 36-108	*Clostridium citroniae* WAL-17108	*Clostridium termitidis* CT1112
Bacteroides ovatus ATCC 8483	*Clostridium hylemonae* DSM 15053	*Clostridium thermocellum* BC1
Bacteroides ovatus SD CC 2a	*Clostridium lentocellum* DSM 5427	*Clostridium thermocellum* JW20
Bacteroides ovatus SD CMC 3f	*Clostridium symbiosum* ATCC 14940	*Ethanoligenens harbinense* YUAN-3
Bacteroides ovatus V975	*Eubacterium rectale* DSM 17629	*Faecalibacterium prausnitzii* A2-165
Bacteroides reticulotermitis JCM 10512	*Herbinix luporum* SD1D	*Mageeibacillus indolicus* UPII9-5
Bacteroides salanitronis DSM 18170	*Lachnoclostridium phocaeense* Marseille-P3177	*Ruminiclostridium* sp. KB18
Bacteroides thetaiotaomicron 7330	*Lachnoclostridium phytofermentans* ISDg	*Ruminiclostridium thermocellum* AD2
Bacteroides thetaiotaomicron VPI-5482	*Lachnoclostridium* sp. YL32	*Ruminiclostridium thermocellum* ATCC 27405
Bacteroides vulgatus ATCC 8482	*Roseburia hominis* A2-183	*Ruminiclostridium thermocellum* DSM 1313
Bacteroides xylanisolvens SD CC 1b	*Roseburia inulinivorans* CAG :15	*Ruminiclostridium thermocellum* DSM 2360
Bacteroides xylanisolvens XB1A	*Roseburia inulinivorans* DSM 16841	*Ruminiclostridium thermocellum* YS
Eubacteriaceae	*Ruminococcus torques* L2-14	*Ruminococcaceae bacterium* CPB6
[Eubacterium] eligens ATCC 27750	*Tyzzerella nexilis* DSM 1787	*Ruminococcus albus* 7 = DSM 20455
Acetobacterium woodii DSM 1030		*Ruminococcus bicirculans* 80/3
Eubacterium limosum ATCC 8486		*Ruminococcus bromii* L2-63
Eubacterium limosum KIST612		*Ruminococcus champanellensis* 18P13 = JCM 17042
Eubacterium limosum SA11		*Ruminococcus flavefaciens* 17
Eubacterium plexicaudatum ASF492		
Eubacterium rectale CAG 36		

Tableau B.9 – Liste des organismes présents dans MACADAM (version novembre 2018) pour le genre *Bacteroides*, et les familles *Lachnospiraceae*, *Ruminococcaceae* et *Eubacteriaceae*.

Au contraire, les membres de la famille des *Ruminococcaceae* semblent être plus spécialisés dans la fermentation du pyruvate en lactate et dans la dégradation des polymères végétaux tels que la cellulose. Les *Lachnospiraceae* ont une position intermédiaire partageant les voies de dégradation des glucides et les processus de fermentation avec les *Bacteroides* (melibiose degradation, acetate formation from acetyl CoA, et xylose degradation) ou avec les *Eubacteriaceae* (fructose degradation pathway). Les bactéries *Eubacteriaceae* présentent certaines spécificités métaboliques telles que l'absence de certaines voies (D-arabinose degradation I, lactose and galactose degradation I, melibiose degradation et (1,4)-β-xylan degradation) et la présence de quatre voies de dégradation du pyruvate.

Ces résultats mettent en avant certaines spécificités métaboliques de ces quatre taxons mais ils doivent être analysés avec prudence car chacun de ces taxons n'est décrit que par un nombre limité d'organismes dans MACADAM (Tableau B.12, page 123). Par ailleurs, entre la date de parution de l'article Read et al. (2019) et l'écriture de ce manuscrit nous avons noté que la famille des *Ruminococcaceae* avait subi un profond remaniement. En effet, selon Zhang et al. (2018), une partie des genres composant cette famille forme un groupe monophylétique dans l'ordre des *Clostridiales*. Une nouvelle famille, les *Hungateiclostridiaceae* a été proposée et retenue (Oren et Garrity 2018a). Le tableau B.10 présente les nouvelles combinaisons des noms d'espèces ainsi que leur famille d'appartenance. Dans le cas de notre étude, douze organismes qui appartenaient a la famille *Ruminococcaceae* appartiennent désormais à la famille des *Hungateiclostridiaceae* soit 43% des organismes appartenant à la famille des *Ruminococcaceae* dans MACADAM version novembre 2018. Ceci montre la fragilité des résultats apportées à partir d'une inférence fonctionnelle sur des rangs taxonomiques élevées.

B.3.3 Analyse du potentiel fonctionnel des genres *Blautia*, *Ruminococcus* et *Mediterraneibacter*

B.3.3.1 Contexte

Dans la partie B.1, page 79, nous avons précisé l'affiliation taxonomique des espèces métagénomiques ainsi que de génomes RefSeq. Ce travail a été effectué à partir de l'analyse d'un arbre phylogénétique reconstruit à partir de protéines marqueurs et par le calcul de l'indice ANI. L'arbre phylogénétique (Figure B.II) a permis de mettre en évidence des regroupements d'espèces métagénomiques selon quatre groupes affiliés à quatre genres : *Blautia*, *Mediterraneibacter*, *Megasphaera* et *Ruminococcus*.

Lachnospiraceae	**Nouvelle combinaison**	
Clostridium lentocellum	*Cellulosilyticum lentocellum*	*Lachnospiraceae*
Ruminococcaceae	**Nouvelle combinaison**	**Nouvelle famille**
[Clostridium] cellulolyticum	*Ruminiclostridium cellulolyticum*	*Hungateiclostridiaceae*
[Clostridium] clariflavum	*Hungateiclostridium clariflavum*	
[Clostridium] stercorarium	*Thermoclostridium stercorarium*	
Acetivibrio cellulolyticus	*Hungateiclostridium cellulolyticum*	
Clostridium papyrosolvens	*Ruminiclostridium papyrosolvens*	
Clostridium termitidis	*Ruminiclostridium cellobioparum**	
Clostridium thermocellum	*Hungateiclostridium thermocellum*	
Mageeibacillus indolicus	*Mageeibacillus indolicus*	
Ruminiclostridium sp. KB18	*Hungateiclostridiaceae bacterium* KB18	
Ruminiclostridium thermocellum	*Hungateiclostridium thermocellum*	

Tableau B.10 – Nouvelles combinaisons de noms d'espèces et leur nouvelle famille d'appartenance par rapport à la parution de l'article Read et al. (2019). Les noms de souche ne sont pas indiqués hormis pour *Ruminiclostridium* sp. KB18 . * : *Clostridium termitidis* est devenu : *Ruminiclostridium cellobioparum* subsp. termitidis.

B.3.3.2 Problématique

Les espèces appartenant aux quatre genres ayant permis l'affiliation taxonomique des espèces métagénomiques (Partie B.1 page 79) peuvent elles être discriminées sur la base de leur potentiel métabolique ?

B.3.3.3 Matériel et méthodes

Nous avons utilisé MACADAM afin d'obtenir l'information fonctionnelle disponible pour les genres *Blautia*, *Mediterraneibacter* et *Ruminococcus*. Ces analyses ont été effectuées avec la version février 2019 de MACADAM. Depuis, certaines espèces du genre *Ruminococcus* ont été transférées dans le genre *Mediterraneibacter* (Togo et al. 2018 ; Oren et Garrity 2019a). Nous avons adapté les résultats en conséquence. Nous avons ajouté également *Megasphaera stantonii* AJH120, afin d'obtenir des informations fonctionnelles sur la kSGB 1. A ces trois genres, nous avons ajouté l'information fonctionnelle de *Clostridioides difficile* 630, cet organisme étant l'organisme extérieur aux genres présents dans l'arbre phylogénétique. Nous avons seulement considéré la souche 630 de *Clostridioides difficile*.

Les scores de complétude (PS) de chaque voie métabolique de chaque organisme inclus dans MACADAM sont ensuite réunis et représentés sous la forme d'une carte thermique hiérarchique (de l'anglais : *heatmap*). Pour cela, nous avons utilisé R et la librairie *gplots* contenant la fonction *heatmap.2*. Celle-ci utilise comme méthode de regroupement le saut maximum (en anglais : *complete linkage methods*) et comme métrique, la distance euclidienne.

B.3.3.4 Résultats et Discussion

Genre	Espèce et souche
Blautia	*B. hansenii* DSM 20583
	B. sp. N6H1-15
	B. sp. YL58
Clostridioides	*C. difficile* 630
Megasphaera	*M. stantonii* AJH120
Ruminococcus	*R. albus* 7 = DSM 20455
	R. bicirculans 80/3
	R. bromii L2-63
	R. champanellensis 18P13 = JCM 17042
	R. flavefaciens 17
	Mediterraneibacter torques L2-14 (*Ruminococcus torques* L2-14)*

Tableau B.11 – Espèces bactériennes présentes dans MACADAM (version février 2019) pour les genres *Blautia*, *Mediterraneibacter* et *Ruminococcus* ainsi que les souches *Megasphaera stantonii* AJH120 et *Clostridioides difficile* 630. * : cette espèce est en cours de transfert vers un autre genre. En effet *[Ruminococcus] torques* L2-14 appartient au genre *Mediterraneibacter*, mais ce changement n'est pas encore publié de manière valide et répercuté dans le NCBI Taxonomy utilisé dans cette version de MACADAM. Par souci de clarté, nous utiliserons le nom *Mediterraneibacter torques* L2-14 dans ces résultats.

MACADAM Les espèces présentes dans MACADAM sont listées dans le tableau B.11. Le tableau B.13 indique la couverture au niveau de l'espèce de ces trois genres. Ainsi, MACADAM ne contient que 6%, 25% et 50% des espèces du genre *Blautia*, *Mediterraneibacter* et *Ruminococcus* respectivement.

Le genre *Blautia* est représenté dans MACADAM par trois souches, dont la souche type de l'espèce : *B. hansenii*. L'espèce type de ce genre (*Blautia coccoides*) n'est pas présente dans MACADAM. Les deux autres souches appartiennent au genre *Blautia* mais leurs espèces n'ont pas encore été définies. Le genre *Ruminococcus* est constitué de 5 souches dans MACADAM, chacune étant la souche type des différentes espèces. *R. flavefaciens* est l'espèce type du genre *Ruminococcus*. Concernant le genre *Mediterraneibacter*, une seule souche est présente dans MACADAM et de manière indirecte. Bien que le genre *Mediterraneibacter* soit présent dans MACADAM, aucune information fonctionnelle ne lui est associée. Une souche appartenant à l'espèce *Ruminococcus torques* est néanmoins présente lors de l'interrogation de MACADAM

sur le genre *Ruminococcus*. Cette souche est actuellement en cours de transfert vers le genre *Mediterraneibacter* (Togo et al. 2018) et ce changement n'est pas encore répercuté sur les données RefSeq et les données taxonomiques utilisées pour générer MACADAM version février 2019. Il ne s'agit pas de l'espèce type du genre *Mediterraneibacter*, ni de la souche type de l'espèce *Ruminococcus torques* (celles-ci étant respectivement *Mediterraneibacter massiliensis* et *Ruminococcus torques* ATCC 27756). Par souci de clarté et bien que ce changement n'a pas été publié en même temps que le genre *Mediterraneibacter* (Oren et Garrity 2019a), nous nommerons cette souche *Mediterraneibacter torques* L2-14 durant ce travail. Les souches *Clostridioides difficile* 630 et *Megasphaera stantonii* AJH120 (souche type de l'espèce *Megasphaera stantonii*) sont présentes dans MACADAM. Les origines ainsi que le nombre de voies associées à chacun des organismes sont présentés dans le tableau B.12.

Souche	**Origine et qualité**	**Voie**
B. hansenii DSM 20583	RefSeq (Complete genome)	94
B. sp. N6H1-15	RefSeq (Complete genome)	97
B. sp. YL58	RefSeq (Complete genome)	114
C. difficile 630	MicroCyc	236
Megasphaera stantonii AJH120	RefSeq (Complete genome)	91
R. albus 7 = DSM 20455	RefSeq (Representative genome)	108
R. bicirculans 80/3	RefSeq (Complete genome)	81
R. bromii L2-63	MicroCyc	153
R. champanellensis 18P13	MicroCyc	174
R. flavefaciens 17	MicroCyc	166
Mediterraneibacter torques L2-14	MicroCyc	212

Tableau B.12 – Origine et qualité des génomes servant de base aux PGDBs présentes dans MACADAM pour les souches d'intérêt. Le nombre de voies est également présenté. *R. champanellensis* 18P13 = JCM 17042 a été abrégé en *R. champanellensis* 18P13.

Les résultats de l'analyse fonctionnelle sont présentés figure B.XI. Les 4 groupes d'organismes de l'arbre phylogénétique (Figure B.II, page 89) ne sont pas observés lors de l'analyse fonctionnelle avec MACADAM. *Clostridioides difficile* 630, organisme servant de groupe extérieur à notre arbre phylogénétique, ne retrouve pas sa position distante des autres souches et est inclus dans l'une des deux branches majeures de nos résultats. Il en va de même pour *Megasphaera stantonii* AJH120. Les différentes souches composant le genre *Ruminococcus* ne forment pas un groupe unique. A l'inverse, les organismes

du genre *Blautia* présentent une certaine proximité entre eux. *Mediterraneibacter torques* L2-14 est quant à lui proche de *Ruminococcus flavefaciens* 17 et *Ruminococcus bromii* L2-63, ce qui est en contradiction avec l'arbre phylogénétique, le genre *Mediterraneibacter* étant plus proche du genre *Blautia* que du genre *Ruminococcus*.

Comment peut-on expliquer l'absence de consensus entre le regroupement selon la taxonomie et celui en fonction du potentiel métabolique ? Plusieurs hypothèses peuvent être avancées : (i) la première concerne la présence de faux négatif : l'identification des voies métaboliques repose sur la qualité des PGDBs construites (dans MACADAM via Pathway Tools) ou récupérées à partir de MicroCyc dans MACADAM. La qualité de ces PGDBs dépend de la qualité d'annotation de ces génomes, celle-ci découlant de la qualité d'assemblage de ces mêmes génomes. Dans MACADAM nous avons fait le choix de construire des PGDBs uniquement à partir de génomes de haute qualité labellisés « Complete genome » issus de la base RefSeq. D'autre part, afin de disposer des annotations les plus complètes et bénéficier du travail de curation par des experts, nous avons ajouté la collection de PGDBs de MicroCyc. Toutefois, en dépit de ces précautions, l'absence de certaines voies métaboliques pour certains organismes constitue un faux négatif et s'explique vraisemblablement par une absence d'annotation plutôt que par un déficit de la fonction. En effet, dans le tableau B.12, les souches dont les PGDBs sont issues de MicroCyc présentent le nombre de voies métaboliques le plus élevé. (ii) La seconde hypothèse concerne l'effort d'annotation des génomes pour certains organismes. En effet, le nombre plus élevé de voies métaboliques identifiées dans *Clostridioides difficile* 630 résulte d'un effort d'annotation manuelle plus important car considéré comme génome de référence dans RefSeq (Sebaihia et al. 2006 ; Larocque, Chénard et Najmanovich 2014). Cet effort est lié à l'intérêt scientifique pour cette espèce bactérienne. En effet, l'espèce bactérienne *Clostridioides difficile* est la principale responsable d'infection nosocomiale (Czepiel et al. 2019). (iii) Une troisième hypothèse pour expliquer cette absence de consensus entre taxonomie et potentiel fonctionnel est en lien avec la redondance fonctionnelle. Les quatre genres bactériens sont retrouvés dans le même type d'habitat à savoir le tractus digestif anaérobie. Pour s'établir, perdurer et se développer, ces bactéries partagent des aptitudes fonctionnelles ou phénotypes proches et seules la comparaison de leur phylogénie, à partir de protéines marqueurs de leur histoire évolutive, permet de les discriminer.

FIGURE B.XI – Classification hiérarchique des organismes en fonction du score PS (Pathway Score) des voies métaboliques identifiées dans leur génome. La couleur varie en fonction du Pathway Score des voies (Partie B.2.1, page 101). Un score de 0 (vert) indique une voie dont aucune réaction n'est annotée dans les génomes et qui est donc absente, tandis qu'un score de 1 (rouge) indique une voie dont l'ensemble des réactions est annoté.

B.3.4 Conclusion

Nous avons utilisé MACADAMExplore afin de comparer le potentiel fonctionnel de taxons sur la base du calcul du PS. Nous avons ainsi pu mettre en évidence des spécificités fonctionnelles dans le cadre des résultats que nous avons produits pour l'étude de Read et al. (2019). Par ailleurs nous avons exploré le potentiel fonctionnel des quatre genres ayant permis l'affiliation taxonomique d'espèces métagénomiques (Partie B.1, page 79). L'ensemble de ces résultats soulève certaines limites de l'analyse du potentiel fonctionnel des procaryotes à partir de leur données génomiques.

Une première limite est celle liée à la présence de faux négatifs résultant d'un travail d'annotation de génome insuffisant ou erroné. Ceci découle du nombre limité de génomes de qualité « reference genome » : qui associe haute qualité d'assemblage et des annotations fonctionnelles manuelles. Actuellement dans MACADAM, 118 PGDBs sont générées à partir de « reference genome », les 11 676 restantes étant construites à partir de « complete genome ».

Une seconde limite concerne la couverture limitée de la diversité procaryotique pour chacun des taxons d'intérêts (Tableau B.13). La diversité restreinte de MACADAM s'explique par le nombre limité de génomes de qualité « complete genome » dans la base de données RefSeq ou MycroCyc pour ces taxons.

Taxon	**Rang taxonomique**	**% dans MACADAM**
Lachnospiraceae	Famille	8,5% des genres
Ruminococcaceae	Famille	10% des genres
Eubacteriaceae	Famille	18% des genres
Bacteroides	Genre	19% des espèces
Blautia	Genre	6% des espèces
Mediterraneibacter	Genre	25% des espèces
Ruminococcus	Genre	50% des espèces

Tableau B.13 – Pourcentages de genre avec au moins une souche présente dans MACADAM dans les familles *Lachnospiraceae*, *Ruminococcaceae*, *Eubacteriaceae* ainsi que le pourcentage d'espèces avec au moins une souche présente dans les genres *Bacteroides*, *Blautia*, *Mediterraneibacter* et *Ruminococcus*. D'après le NCBI Taxonomy le 15/09/2019.

Une troisième limite est relative à la variabilité de la qualité d'annotation des génomes utilisés par MACADAM qui peut générer un biais d'interprétation du potentiel fonctionnel. Ce biais est lié à la « surconnaissance » de

certains taxons qui ont bénéficié d'un effort particulier de recherche de la part de la communauté scientifique comparativement à d'autres ayant reçu moins d'attention.

B.4 Analyses comparées de l'inférence des voies métaboliques d'une communauté artificielle à partir de MACADAM et de données métagénomiques

B.4.1 Introduction

MACADAM est une base de données de voies métaboliques et de caractéristiques fonctionnelles reliées à une information taxonomique. L'objectif de cette partie est de confronter les voies métaboliques associées à certains taxons dans MACADAM aux voies métaboliques qu'il est possible d'obtenir lors de l'analyse de lectures issues du séquençage métagénomique d'un milieu contenant ces mêmes taxons. Nous présentons dans un premier temps le jeu de données utilisé et dans un second temps les deux chaînes de traitement : l'une utilisant MACADAMExplore (Figure B.XII **A**) et l'autre permettant le traitement de lectures issues de séquençage métagénomique (Figure B.XII **B**) : le pipeline HUMAnN2 (Franzosa et al. 2018). Les résultats issus de ces deux chaînes de traitement seront comparés et discutés.

B.4.2 Matériels et Méthodes

B.4.2.1 Données d'intérêt

Les données de séquençage métagénomiques sont décrites dans l'article de Jovel et al. (2016). Les lectures issues du séquençage de ces 31 communautés sont disponibles librement sur le portail Sequence Read Archive (SRA, Leinonen, Sugawara et Shumway 2011) avec l'identifiant SRP059928.

Afin de comparer les résultats d'une analyse fonctionnelle avec MACADAM avec ceux issus d'un séquençage métagénomique, nous avons choisi un jeu de données où les espèces bactériennes sont connues et leurs génomes séquencés, annotés et déposés dans les banques de données génomiques. Le jeu de données nommé MIX1 (identifiant SRA : SRR2081071) remplit l'ensemble de ces critères. Ce jeu de données correspond à une communauté bactérienne artificielle composée d'un mélange de 11 espèces bactériennes appartenant à 7 genres (Tableau B.14). Ces 11 espèces sont cultivées dans des conditions de laboratoire standard.

Genre	Espèce bactérienne
Bacteroides	*B. thetaiotaomicron*
	B. vulgatus
Bifidobacterium	*B. animalis*
	B. breve
Escherichia	*E. coli*
Lactobacillus	*L. delbrueckii*
	L. helveticus
	L. plantarum
Paeniclostridium	*P. sordellii*
Salmonella	*S. enterica*
Streptococcus	*S. pyogenes*

Tableau B.14 – Espèces bactériennes composant la communauté bactérienne MIX1.

B.4.2.2 Pipeline MACADAMExplore

Pour obtenir les informations fonctionnelles de la communauté bactérienne d'intérêt en utilisant MACADAMExplore (Figure B.XII **A**) nous récupérons les noms binomiaux de chaque espèce bactérienne présente dans le MIX1 et utilisons le script MACADAMExplore sans contrainte de score. Le nombre de PGDBs disponibles pour chaque espèce bactérienne dans MACADAM est présenté dans le tableau B.15. Ces PGDBs ont été obtenus soit à partir de : (i) génomes provenant de RefSeq avec un niveau d'assemblage « complete genome » (Page 105, paragraphe « Building PGDBs between RefSeq and MetaCyc » de l'article MACADAM), (ii) MicroCyc, une collection de PGDBs ayant été soumis à un processus de curation manuelle (Page 106, paragraphe « Embedding PGDBs from MicroCyc »).

Nous avons ensuite développé un script python permettant de rassembler l'ensemble des résultats obtenus par MACADAMExplore en un seul fichier et permettant de filtrer sur le score de complétude (Pathway Score) et le score de fréquence (Pathway Frequency Score) de chacune des voies métaboliques. La signification de ces scores est expliquée dans la figure 2, page 107 de l'article de MACADAM.

Nous avons développé un script python permettant d'extraire l'intégralité des voies métaboliques présentes dans chacune des espèces, le nombre de PGDBs dans lesquelles les voies métaboliques sont présentes ainsi que les valeurs de leur Pathway Score.

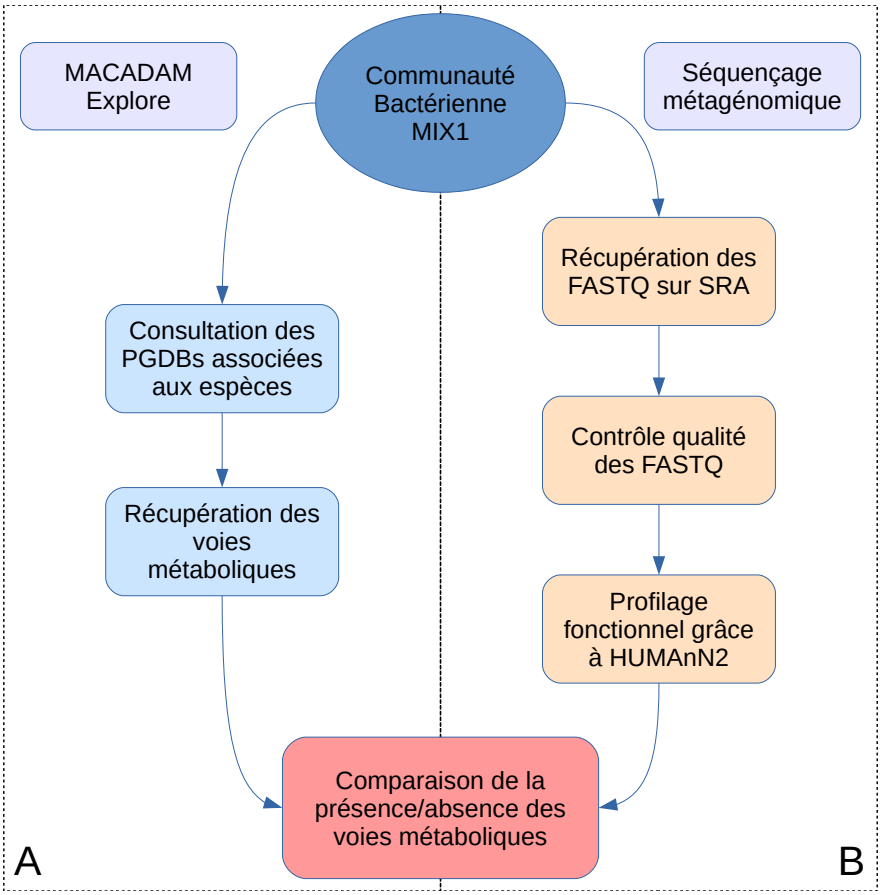


FIGURE B.XII – Schéma des deux chaînes de traitement d'analyse permettant la comparaison des voies métaboliques présentes : **A** : MACADAMExplore ; **B** : Séquençage métagénomique en utilisant HUMAnN2.

B.4.2.3 Pipeline de traitement des données métagénomiques

Contrôle qualité Le fichier de données contenant les lectures issues du séquençage de MIX1 a été extrait de SRA et converti en FASTQ grâce au SRA Toolkit (`https://github.com/ncbi/sra-tools`) mis à disposition par le NCBI.

Nous avons, dans un premier temps, contrôlé la qualité des séquences grâce à l'outil FASTQC (ANDREWS 2010), disponible sur la plateforme Genotoul Bioinfo (version 0.11.7). FASTQC permet d'évaluer le score de qualité Phred moyen en fonction de la position du nucléotide dans la séquence. Le score Phred traduit la probabilité d'erreur d'identification d'une nucléotide par le séquenceur. Il est basé sur une échelle logarithmique. Nous avons vé-

Espèce bactérienne	PGDBs MACADAM	ChocoPhlAn
Bifidobacterium animalis	22	15 (13)
Bifidobacterium breve	41	7(2)
Bacteroides vulgatus	1	4 (1)
Bacteroides thetaiotaomicron	2	2 (1)
Paeniclostridium sordellii	1	2 (0)
Lactobacillus plantarum	83	16 (6)
Lactobacillus delbrueckii	20	11 (4)
Lactobacillus helveticus	17	7 (4)
Escherichia coli	667	764 (57)
Streptococcus pyogenes	119	46 (19)
Salmonella enterica	565	520 (40)

Tableau B.15 – Information disponible pour chaque espèce bactérienne présente dans MIX1 dans la base de données MACADAM (nombre de PGDBs) et dans la base de données ChocoPhlAn (nombre d'assemblage génomique). Le nombre de génomes de qualité « complete genome » composant la base de données ChocoPhlAn est indiqué entre parenthèses.

rifié que toutes les positions des séquences présentaient en moyenne un score Phred de 20 minimum, ce score représentant une fiabilité a 99% lors de l'identification du nucléotide durant le séquençage. Ce seuil est couramment utilisé pour valider la qualité des lectures.

HUMAnN2 HUMAnN2 (The HMP Unified Metabolic Analysis Network 2, Franzosa et al. 2018) est un pipeline d'analyse de lectures issues de séquençage métagénomique permettant d'identifier les micro-organismes présents dans un milieu et d'obtenir un profil de présence/absence de voies métaboliques. HUMAnN (Abubucker et al. 2012) a été développé pour l'étude des profils fonctionnels des métagénomes produits dans le cadre du « Human Microbiome Project » (HMP, Methé et al. 2012). HUMAnN2, seconde version du pipeline, a été développé afin de lier une information taxonomique aux profils fonctionnels, de ne pas se limiter aux organismes présents dans le projet HMP et de limiter le recours à la recherche par des séquences nucléiques traduites en acides aminés afin d'optimiser le temps d'exécution du pipeline. Nous avons retenu HUMAnN2 car les résultats d'analyse sont comparables à ceux issus de MACADAMExplore. Pour ces deux pipelines, les voies métaboliques sont construites à partir de MetaCyc. Nous avons installé HUMAnN2 sur la plateforme Genotoul Bioinfo afin de bénéficier de sa puissance de calcul et utilisé le pipeline sur les données MIX1.

HUMAnN2 est composé de quatre modules incluant 4 outils (Figure B.XIII). Dans le premier module du pipeline (**A**) le logiciel MetaPhlan2 (Metagenomic Phylogenetic Analysis, Truong et al. 2015) permet d'associer chaque lecture à un clade d'organismes associé à une taxonomie grâce à l'utilisation de gènes marqueurs spécifiques a certains clades. Si une lecture est reconnue comme faisant partie d'une taxonomie alors elle est liée au profil fonctionnel de cet organisme (**B**). Si la lecture n'a pas été associée à une taxonomie, alors celle-ci est alignée contre une banque de données protéiques (**C**). Les résultats des deux dernières étapes sont ensuite agrégés afin de reconstruire les voies métaboliques présentes dans le métagénome (**D**).

Etape A : assignation taxonomique à partir d'une base de données de gènes marqueurs. Avant d'utiliser HUMAnN2, il est nécessaire de s'assurer que les lectures passent un contrôle qualité, comme expliqué précédemment. Il est important de noter que HUMAnN2 ne prend pas en compte les fichiers de lectures séquençées en paire. En effet, lors de l'étape d'alignement des séquences, il est fréquent que la première lecture corresponde à une séquence codante, tandis que la seconde lecture associée ne corresponde pas à cette même séquence codante. Dans ce cas, l'alignement de séquence n'aboutirait pas et la paire de lectures ne pourrait être associée à une protéine. Par ailleurs, dans un séquençage par paire, la seconde lecture présente également un score de qualité plus faible que la première paire. Ces raisons justifient l'utilisation de la première lecture de la paire séquencée dans le pipeline HUMAnN2.

Afin d'associer une lecture à un clade bactérien, l'outil MetaPhlan2 interroge une base de données qui contient plus d'un million de séquences de gènes marqueurs spécifiques à un clade bactérien, archées, virus et eucaryotes. Ces gènes sont issus de génomes déposés dans « The Integrated Microbial Genomes & Microbiomes » (IMG/M, Chen et al. 2019) dont les gènes marqueurs ont été identifiés selon la méthode présentée dans Segata et Huttenhower (2011). Si une lecture est associée à une taxonomie par MetaPhlan2 et que l'abondance relative des lectures appartenant à cette espèce dans l'échantillon est supérieure à 0.01% alors le nom de l'espèce est mis en mémoire pour la prochaine étape de HUMAnN2.

Etape B : alignement sur une base de données pangénomique HUMAnN2 fait ensuite appel à une collection de pangénomes (Figure B.XIII **B**) : ChocoPhlAn. Le pangénome (Tettelin et al. 2005) est composé : (i) de l'ensemble non-redondant des gènes présents dans toutes les souches bactériennes d'une même espèce bactérienne, ainsi que (ii) de l'ensemble

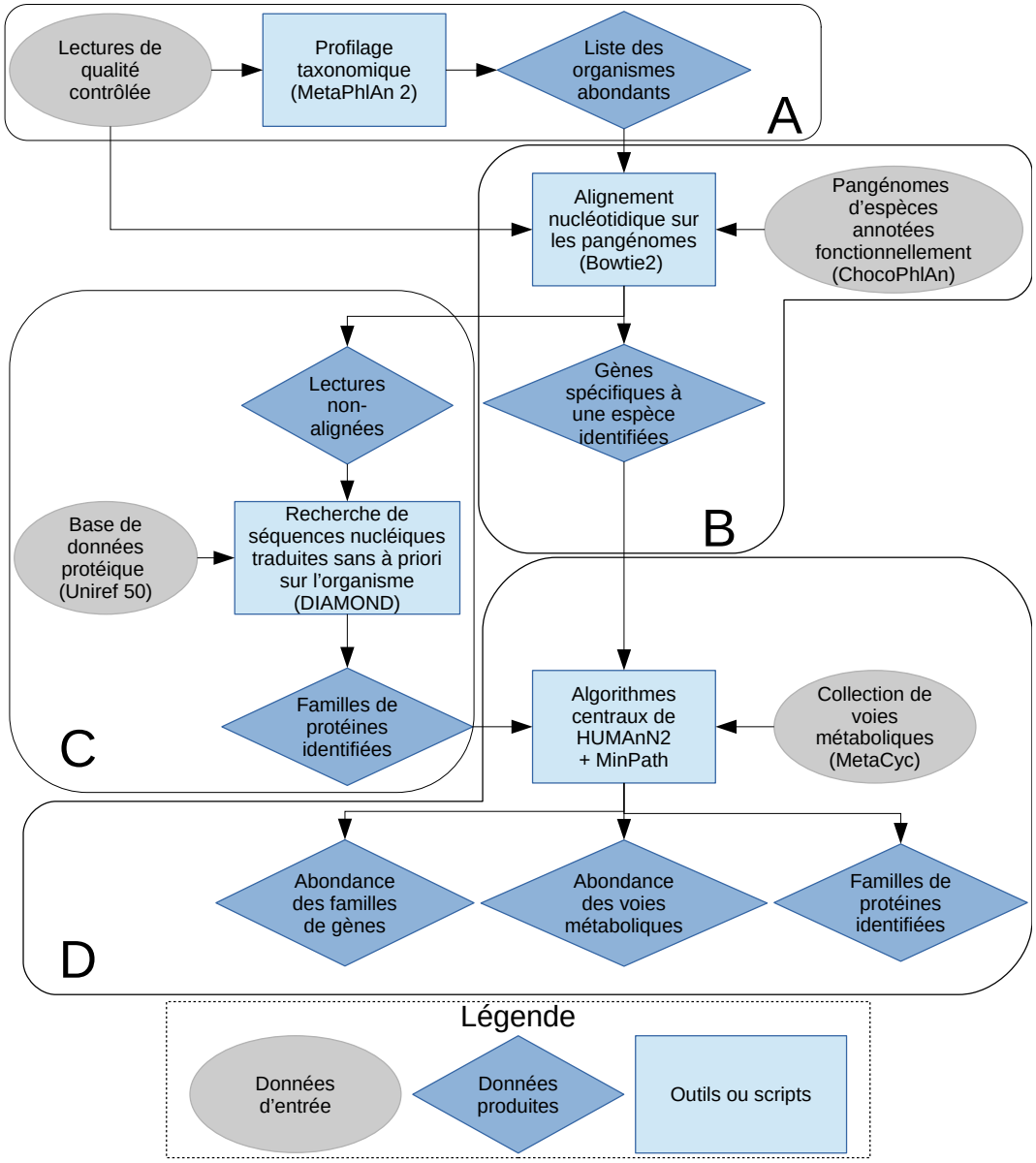


FIGURE B.XIII – HUMAnN2 pipeline. Composé de quatre modules : **A** : analyse phylogénétique des lectures issues de séquençage métagénomique par MetaPhlAn v2.0 ; **B** : alignement sur une base de données de pangénomes ; **C** : recherche par des séquences nucléiques traduites contre une base de données protéiques ; **D** : reconstruction des voies métaboliques.

D'après `http://huttenhower.sph.harvard.edu/humann2`

non redondant des gènes propres aux écotypes et aux souches bactériennes. ChocoPhlAn rassemble l'intégralité des génomes de GenBank et l'intégralité des annotations des séquences codantes de ces mêmes génomes. Le nombre d'assemblages génomiques disponibles dans la base de données ChocoPhlAn pour les 11 bactéries de la communauté est indiqué dans le tableau B.15. PhyloPhlAn (Segata et al. 2013) est utilisé afin de valider l'affiliation taxonomique des génomes. Les lectures correspondant à des séquences codantes sont ensuite regroupées au seuil de 97% d'identité grâce a UCLUST (Edgar 2010). Une seule séquence appelée centroïde est conservée. Ce centroïde est ensuite traduit en une séquence d'acides aminés qui est recherchée dans une banque de données de protéines de références : UniRef90 (Suzek et al. 2015) ou dans UniRef50 si absente dans la première.

Lors de l'étape d'assignation taxonomique (étape **A**), les noms d'espèces bactériennes identifiées et suffisamment abondantes parmi les lectures ont été mis en mémoire. Les pangénomes présents dans ChocoPhlAn sont alors concaténés afin de générer un index Bowtie 2 (Langmead et Salzberg 2012). L'intégralité des lectures est ensuite aligné contre cet index. Si une lecture s'aligne, alors celle-ci est associée avec l'annotation protéique UniRef de la séquence présente dans ChocoPhlAn.

Étape C : Recherche par des séquences nucléiques traduites Les lectures qui n'ont pas pu être alignées lors de l'étape précédente sont alignées grâce à l'outil DIAMOND (Double index alignment of next-generation sequencing data) (Buchfink, Xie et Huson 2015) contre la banque de données UniRef90. Cette deuxième étape permet d'associer les lectures nucléiques à des protéines.

Étape D : Reconstruction des voies métaboliques La dernière étape du pipeline lie les annotations UniRef obtenues lors des parties **B** et **C** aux réactions MetaCyc et, à partir de ces mêmes réactions, forme des voies métaboliques présentes dans MetaCyc. Cette étape est réalisée grâce à MinPath (Ye et Doak 2009)

Fichiers de sortie HUMAnN2 produit trois fichiers de sortie : le premier présente l'abondance des familles de gènes dans la communauté et dans chaque espèce reconnue lors de l'étape **A**. Les abondances sont exprimées en lectures par kilobase (RPK) . Le deuxième fichier contient l'abondance des voies métaboliques reconstruites dans la communauté globale, puis en détail dans chaque espèce détectée. L'abondance d'une voie métabolique correspond dans HUMAnN2 au nombre de copies complètes de voies métaboliques

présentes dans la communauté ou dans l'espèce bactérienne. Le dernier fichier contient la couverture des voies métaboliques dans la communauté globale puis dans les espèces bactériennes détectées. Il s'agit ici d'un score oscillant entre zéro (aucune réaction n'est détectée pour effectuer la voie métabolique) et un (l'intégralité des réactions est détectée dans la communauté pour effectuer la voie métabolique).

B.4.2.4 Comparaison des résultats issus de MACADAM et du pipeline d'analyse de données métagénomiques

Pour comparer les résultats obtenus par MACADAM et HUMAnN2, nous avons développé un script python disponible à l'adresse : `https://github.com/maloleboulch/MACADAMvsHUMAnN2`. Les fichiers de sortie de HUMAnN2 et les résultats de MACADAM sont utilisés afin de déterminer l'ensemble des voies métaboliques communes et propres à chacune des chaînes de traitement. Les voies présentes dans les deux résultats sont comparées afin de déterminer si celles-ci sont présentes dans les mêmes espèces ou des espèces différentes. Les différents scores associés par MACADAM ou HUMAnN2 sont conservés dans les résultats. Pour évaluer le degré de similarité nucléotidique entre deux génomes, le score ANIm est calculé en utilisant le logiciel pyani (Pritchard et al. 2015).

B.4.3 Résultats & Discussion

B.4.3.1 Contrôle qualité des lectures métagénomiques

Grâce à FASTQC, on observe que le fichier de lecture utilisé (MIX1) ne contient pas de trace des amorces utilisées lors du séquençage. Le score de qualité Phred des séquences ainsi que celui par base en fonction de la position de celle-ci (Figure B.XIV) est supérieur à 28, ce qui suggère que les fichiers déposés dans SRA ont déjà été pré-traités en écartant les séquences de faible qualité. Dans la figure B.XIV, la baisse de qualité des bases vers l'extrémité 3' des séquences s'explique par les limites du séquençage Illumina, en particulier dû à la désynchronisation de l'élongation des lectures au fur et à mesure des cycles (Fuller et al. 2009) conduisant à une accumulation d'erreurs et *in fine* à une baisse du score de qualité. Nous avons choisi les premières paires de séquences issues du séquençage car les secondes paires présentaient des scores PHRED inférieurs à 20.

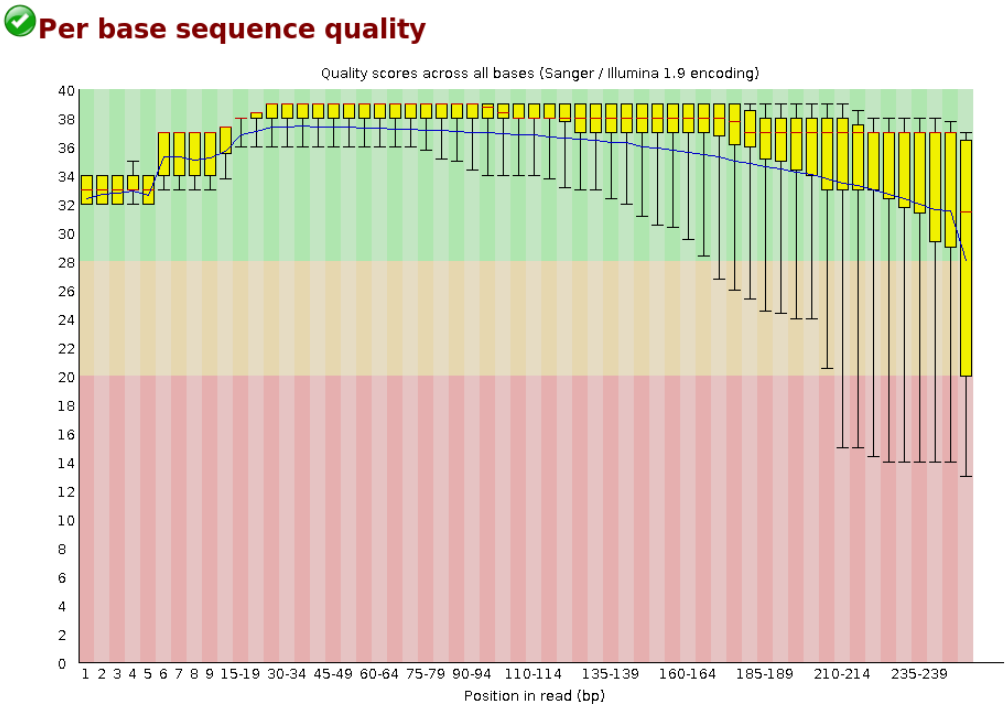


FIGURE B.XIV – Score Phred des bases en fonction de leur position dans la séquence.

B.4.3.2 Analyse des résultats de HUMAnN2

Les 11 espèces bactériennes composant la communauté sont toutes détectées lors de l'étape **A** d'assignement taxonomique de l'analyse des données métagénomiques par le pipeline HUMAnN2. Par conséquent, les voies métaboliques pour ces espèces sont construites à partir des annotations fonctionnelles des assemblages génomiques présents dans la base de données ChocoPhlAn (Tableau B.15, page 131).

Dans les sorties de HUMAnN2 *Paeniclostridium sordellii* est présent sous le nom de *Clostridium sordellii*. Cela s'explique par la proposition d'un nouveau genre par SASI JYOTHSNA et al. (2016), *Paeniclostridium*, suite à l'isolement d'une nouvelle espèce (*Paraclostridium benzoelyticum*). *Clostridium sordellii* partage avec *Paraclostridium benzoelyticum* de plus nombreuses similitudes biochimiques et phylogénétiques qu'avec les espèces du genre *Clostridium*. *Clostridium sordellii* a donc fait l'objet d'un reclassement dans le genre *Paeniclostridium*. MACADAM se basant sur le NCBI Taxonomy, la nouvelle combinaison espèce-genre est déjà présente. ChocoPhlAn contient encore la combinaison obsolète. Nous considérerons donc *Clostridium sordellii* comme *Paeniclostridium sordellii* dans la suite de nos résultats.

HUMAnN2 détecte une espèce supplémentaire qui n'est pas présente dans la communauté initiale des 11 bactéries : *Clostridium bifermentans*, récemment renommée *Paraclostridium bifermentans* Sasi Jyothsna et al. (2016), suite à l'isolement d'une espèce proche *Paraclostridium benzoelyticum*. Cette dernière est phylogénétiquement plus proche de *Paraclostridium bifermentans*, mais plus éloignée du genre *Clostridium*. Nous avons calculé le score ANI des souches type de l'espèce *Paraclostridium bifermentans* (*Paraclostridium bifermentans* ATCC 638, numéro d'accession NCBI : AVNC00000000.1) et de l'espèce *Paeniclostridium sordellii* (*Paeniclostridium sordellii* ATCC 9714, numéro d'accession NCBI : APWR00000000.1) (Pritchard et al. 2015). Le score ANI entre ces deux souches est de 85.25% de similarité de séquence. La présence de cette espèce dans les résultats de HUMAnN2 peut donc s'expliquer par cette similarité entre *Paraclostridium bifermentans* et *Paeniclostridium sordellii* et la petite taille des lectures analysées.

B.4.3.3 Analyse comparée des résultats des deux pipelines d'analyse

Le nombre de voies métaboliques communes et propres à chaque pipeline d'analyse est présenté à l'aide d'un diagramme de Venn (Figure B.XV). L'utilisation de HUMAnN2 permet d'inférer un total de 286 voies métaboliques dont 135 sont communes à celles inférées par l'utilisation de MACADAM. Parmi ces 135 voies, 16 ne sont pas associées à une espèce par HUMAnN2 alors qu'elles le sont forcément dans MACADAM (puisque la requête est taxonomique). Parmi les 151 voies détectées uniquement par HUMAnN2, 32 ne sont affiliées à aucune espèce bactérienne. La construction de ces voies métaboliques résulte vraisemblablement de l'étape de recherche par des séquences nucléiques traduites effectuée sur les lectures qui n'ont pas pu être alignées contre la base pangénomique ChocoPhlAn (Figure B.XIII, étape **B** et **C**).

Parmi les 714 voies métaboliques extraites à partir de MACADAM, 579 sont propres à MACADAM. Le tableau B.16 détaille, pour chacune des 11 espèces constituant la communauté, le nombre de voies détectées par MACADAM et par HUMAnN2, ainsi que le pourcentage de voies métaboliques détectées par HUMAnN2 qui sont communes à celles proposées par MACADAM. Par exemple, pour l'espèce *Bacteroides thetaiotaomicron* 67% des 28 voies identifiées par HUMAnN2 sont présentes dans MACADAM (Tableau B.16). Pour chaque espèce, le nombre de voies métaboliques inféré par MACADAM est supérieur à celui détecté par HUMAnN2.

Comment expliquer cette différence de nombre de voies inférées en fonction du pipeline utilisé ? Plusieurs hypothèses peuvent être avancées : (i) ces

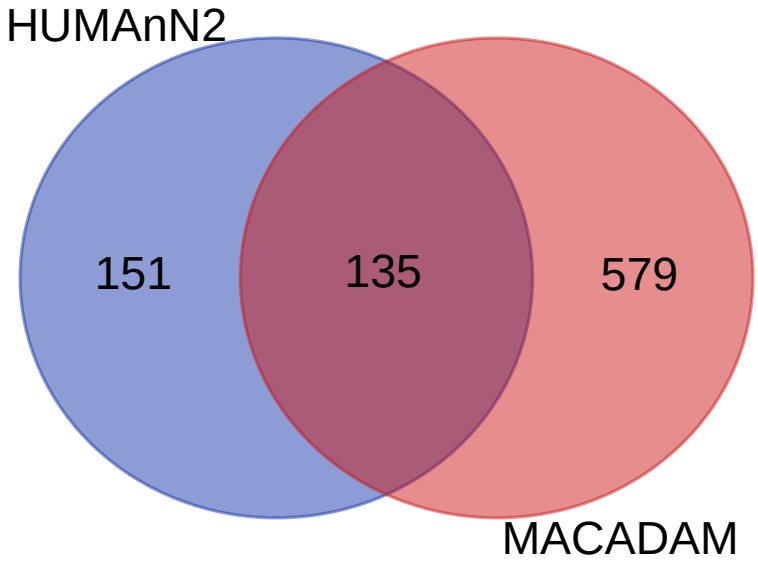


FIGURE B.XV – Nombre de voies métaboliques présentes : uniquement dans les résultats de HUMAnN2, uniquement dans ceux de MACADAM ou communes entre les deux.

Espèce bactérienne	**MACADAM**	**HUMAnN2**	**Voies communes HUMAnN2**
Bacteroides thetaiotaomicron	287	28	67 %
Bacteroides vulgatus	247	22	54 %
Bifidobacterium animalis	211	21	52 %
Bifidobacterium breve	217	18	55 %
Escherichia coli	610	178	50 %
Lactobacillus delbrueckii	195	14	64 %
Lactobacillus helveticus	231	14	64 %
Lactobacillus plantarum	364	23	60 %
Paeniclostridium sordellii	104	12	0 %
Salmonella enterica	498	153	46 %
Streptococcus pyogenes	269	34	61 %

Tableau B.16 – Nombre de voies identifiées dans chaque espèce du MIX1 dans MACADAM et dans HUMAnN2.

deux pipelines disposent chacun de scores qui leur sont propres et permettent de donner une statistique de probabilité de présence de la voie (Pathway Score pour MACADAM, pathway coverage pour HUMAnN2). Dans les résultats présentés, ces scores qui ne sont pas comparables n'ont pas été pris en compte et pourraient réduire le nombre de faux positifs. (ii) Dans MACADAM, l'inférence des voies se fait à partir de génomes complets, tandis que HUMAnN2 la réalise à partir de lectures de 250 paires de bases sans assemblage préalable. (iii) L'identification des voies métaboliques par HUMAnN2 peut être impactée par la profondeur de séquençage : les voies les plus présentes sont les plus détectées. Alors que par MACADAM, la voie est détectée quelle que soit sa redondance dans l'écosystème constitué ici de 11 bactéries. (iv) De plus, comme indiqué dans le tableau B.15, MACADAM présente, pour 8 espèces sur 11, un plus grand nombre de souches pour chaque espèce. Par exemple, *Bifidobacterium breve* est représentée par 41 « complete genomes » dans MACADAM. Dans HUMAnN2, elle n'est représentée que par sept assemblages génomiques, dont deux sont de qualité « complete genome ». Ceci explique un nombre plus grand de voies métaboliques qui correspond à l'union des voies présentes dans ces souches. (v) Le plus grand nombre de voies détecté dans MACADAM peut résulter de la différence de qualité de génome utilisé pour construire les voies. MACADAM, contrairement à ChocoPhlAn, du pipeline HUMAnN2, ne fait appel qu'à des « complete genome » (Tableau B.15, page 131). Hormis pour *Bacteroides vulgatus*, où les deux bases partagent vraisemblablement le même « complete genome », MACADAM contient un nombre toujours supérieur de « complete genome ». Notons que pour *Paeniclostridium sordellii* (Tableau B.16), aucune des voies identifiées par HUMAnN2 n'est identifiée par MACADAM. Pour cet organisme, ChocoPhlAn n'utilise aucun « complete genome » mais a recours à des génomes sous forme de contig (*Paeniclostridium sordellii* VPI 9048 et *Paeniclostridium sordellii* ATCC 9714).

B.4.4 Conclusion

L'avantage majeur des approches métagénomiques réside dans leur indépendance entre l'affiliation taxonomique des lectures qui composent la communauté procaryotique et l'inférence fonctionnelle de cette même communauté procaryotique. L'inférence fonctionnelle à partir de lectures métagénomiques présente l'avantage de capturer la diversité des fonctions *de novo* sans forcément se baser sur des données de référence. Ceci est rendu possible grâce à la recherche de séquences nucléiques traduites. Toutefois cette recherche reste limitée aux protéines répertoriées dans les bases dédiées.

A l'inverse, l'avantage de réaliser une inférence fonctionnelle en utilisant

MACADAM, c'est à dire en partant du nom d'espèce, permet de voir l'intégralité des voies métaboliques pour peu qu'elles soient annotées dans les génomes. Ceci sous-entend une qualité suffisante de génome et une représentation suffisante de la diversité des espèces. Contrairement à l'approche métagénomique, toutes les voies sont observées, quelle que soit leur abondance dans l'écosystème : les rares, comme les abondantes pourvu qu'elles soient reliées à une taxonomie. *A contrario* l'inférence à partir de données métagénomiques est largement dépendante de la profondeur de séquençage. Une profondeur suffisante pouvant entraîner des coûts qui restent importants (Sims et al. 2014). Si l'inférence fonctionnelle à partir de données taxonomiques n'est pas dépendante des technologies de séquençage, elle l'est des erreurs d'affiliation taxonomique des populations qui constituent la communauté. Toutefois, les méthodes d'affiliation taxonomique s'affinent : longtemps restreintes au gène de l'ARNr 16S, l'utilisation de multiples gènes marqueurs pour la reconstruction d'un arbre phylogénétique est maintenant courante (Partie A.1.5.2, page 23, Parks et al. 2018).

Chapitre C

Discussion

Les organismes procaryotes jouent un rôle crucial dans le biota terrestre, car ils catalysent les processus assurant la vie sur Terre et sont les artisans des cycles biogéochimiques (Torsvik, Øvreås et Thingstad 2002). Ces organismes, qui vivent en communauté, ont colonisé la plupart des milieux (Costello et al. 2009 ; Wagner et Loy 2002 ; Fierer, Schimel et Holden 2003). Ces communautés jouent un rôle primaire dans les processus globaux terrestres (Fuhrman 2009), il est donc important d'étudier ces communautés procaryotiques tant au niveau de leur structure qu'au niveau de leurs fonctions. Pour étudier leur structure, des techniques ont été développées permettant d'identifier quels organismes sont présents dans la communauté, quelles sont leurs abondances respectives et quelle est leur dynamique d'évolution (Stubbendieck, Vargas-Bautista et Straight 2016). Le classement des organismes identifiés dans l'arbre du vivant contribue à élargir nos connaissances en microbiologie et permet de réutiliser ces connaissances.

Afin de comprendre et piloter plus finement les processus sous le contrôle de communautés procaryotiques, il convient de déterminer les besoins, les fonctions et l'activité de chaque organisme la composant.

Pour cela, l'accès à l'information génétique permet de parvenir aux fonctions de l'organisme (The Human Microbiome Project Consortium et al. 2012). Mais à défaut d'avoir accès à cette information génétique, il est possible de prédire le potentiel fonctionnel d'un organisme via son identification et sa classification (Langille et al. 2013 ; Asshauer et al. 2015).

Dans ce travail de thèse nous nous sommes penchés sur trois problématiques : (i) la correction d'affiliations taxonomiques d'espèces bactériennes putatives dont les génomes sont reconstruits à partir de données de séquençage métagénomique, (ii) la construction et la mise à disposition d'une base de données permettant de lier un nom taxonomique avec un potentiel fonc-

tionnel. Ce travail est le coeur de la thèse. (iii) Enfin, la confrontation de deux approches permettant d'obtenir la caractérisation fonctionnelle de communautés bactériennes artificielles : via l'information taxonomique disponible sur la communauté, d'une part, et en utilisant les lectures issues de séquençage métagénomique d'autre part.

Affiliation taxonomique d'espèces métagénomiques

Dans la première partie de cette thèse nous nous sommes intéressés à neuf groupes d'espèces métagénomiques obtenus par la méthode du *binning* (Partie A.2.2.2, page 46) dans le cadre d'une étude d'exploration des communautés par séquençage métagénomique chez l'homme présentée dans l'article « Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle », Pasolli et al. (2019). Ces groupes d'espèces métagénomiques ont été affiliés, dans un premier temps, au genre *Blautia* tout en possédant un nom binomial d'espèce les rapprochant du genre *Ruminococcus*. Ceci est incorrect : le premier mot d'une espèce devant être le genre de celle-ci (règle 12a du code de la nomenclature des procaryotes, Parker, Tindall et Garrity 2019). Cette erreur peut être due à la nature polyphylétique du genre *Ruminococcus* (Rainey et Janssen 1995 ; La Reau, Meier-Kolthoff et Suen 2016). En effet, le genre *Blautia* est issu du reclassement d'organismes provenant du genre *Ruminococcus* (Liu et al. 2008). Ces travaux ont été effectués au CIBIO à l'université de Trente (Italie) au sein du laboratoire du professeur Nicola Segata, dans le cadre du parcours doctoral de l'École internationale de recherche d'Agreenium.

Afin de connaître la véritable position de ces neuf groupes d'espèces métagénomiques dans la classification, nous avons reconstruit un arbre phylogénétique en associant aux génomes reconstruits de ces neuf groupes des génomes de références provenant de GenBank. Afin d'affiner ces résultats, nous avons ensuite procédé au calcul d'un indice de parenté génomique : l'ANI (Partie A.1.5.2, page 24).

Nous avons montré que ces neufs groupes d'espèces métagénomiques se répartissent en quatre branches distinctes dans l'arbre phylogénétique reconstruit (Figure B.II, page 89, résumé dans la figure C.I, page 89), contenant respectivement les génomes de référence de : (i) six souches type de *Ruminococcus*, (ii) la souche type d'un genre récemment décrit : *Mediterraneibacter* (Togo et al. 2018), (iii) 10 souches type d'espèces provenant du genre *Blau-*

tia et *Ruminococcus gauvreauii*, (iv) une branche se rapprochant du genre *Megasphaera*.

L'inclusion de génomes de référence dans la reconstruction de l'arbre phylogénétique a permis d'identifier deux groupes comme appartenant à l'espèce *Ruminococcus bicirculans* pour l'un, et au genre *Megasphaera* pour l'autre. Le calcul de l'ANI a confirmé la position du premier groupe, tandis que l'espèce dans le genre *Megasphaera* a été identifiée : *Megasphaera stantonii*.

Les 7 autres groupes voient leur classification précisée au niveau du genre mais non élucidée au niveau de l'espèce. Cette étude met en lumière que la nature polyphylétique du genre *Ruminococcus* est toujours d'actualité, malgré les efforts de reclassement réalisés pour ce genre lors de la définition du genre *Mediterraneibacter*. En effet, *Ruminococcus gauvreauii* ne semble pas appartenir au genre *Ruminococcus*. L'ensemble de ces résultats est présenté dans la figure C.I.

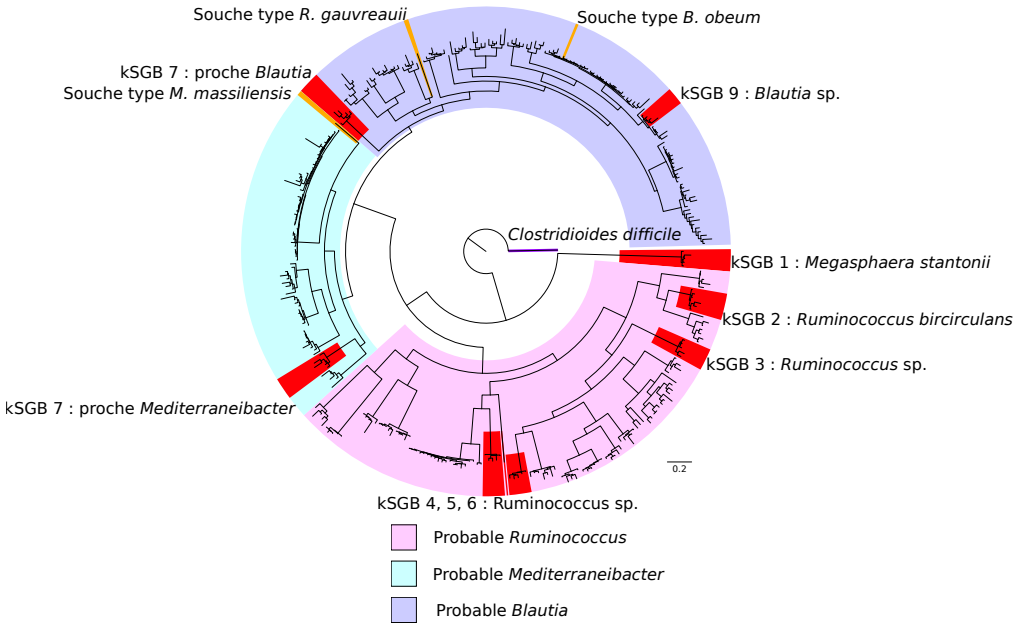


FIGURE C.I – Arbre phylogénétique résumant les résultats mis en évidence dans la partie B.1, page 79.

Dans le cadre de ces travaux, nous mettons en évidence que le pipeline utilisé dans le contexte de l'étude « Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle », PASOLLI et al. (2019), réplique des erreurs produites lors du dépôt de génomes dans GenBank. C'est notamment le cas pour certains des dépôts effectués dans le cadre de l'étude de BROWNE et al.

(2016) (Partie B.1.6, page 99). Dans cette étude les génomes séquencés ont été isolés mais non cultivés avant séquençage. Cette démarche est vraisemblablement à l'origine d'une mauvaise affiliation taxonomique lors du dépôt des séquences dans GenBank. La classification procaryote étant une science en perpétuelle évolution, il est probable que l'erreur d'affiliation résulte des mises à jour successives de la classification non répercutées. Toutefois nous soulignons ici, que lors du dépôt de génomes dans les banques de données génomiques afin de garantir le meilleur positionnement possible, il est conseillé de s'appuyer au moins sur l'ANI (Chun et al. 2018) pour une description d'espèce et sur la reconstruction d'un arbre phylogénétique. Parks et al. (2019) propose une taxonomie se basant sur l'ANI. Dans notre cas, nous avons procédé à une reconstruction d'arbre phylogénétique basée sur de multiples marqueurs, et nous avons complété ces données par un calcul de l'ANI contre les souches types du genre présent dans la banque de données (Beaz-Hidalgo et al. 2015).

Lors de la préparation du protocole expérimental d'une étude, le choix de la base génomique qui va être utilisée est prépondérant. En effet, bien que GenBank contienne un large choix de génomes et donc de diversité génomique permettant d'affilier des génomes inconnus, notre travail présente le fait que certains génomes, issus de séquençage métagénomique, présentent une affiliation fausse. Afin d'éviter ces erreurs, il convient d'utiliser une base de données qui, bien que présentant des organismes moins divers, exclut les génomes provenant d'assemblages métagénomiques tels que RefSeq (Tatusova et al. 2014). Mais ces dernières ne sont pas exemptes de mauvaise affiliation, comme nous avons pu le constater. En effet, dans notre étude sur le genre *Blautia*, des espèces dont les génomes étaient pourtant issus de la base de données RefSeq présentaient une erreur de classification. Ces génomes n'étaient toutefois pas de haute qualité puisqu'il s'agissait de séquences assemblées en scaffold. Le choix de la base de données doit également être guidé par l'étendue de l'étude et de son objectif. Dans le cadre de l'étude de Pasolli et al. (2019), il s'agissait d'un travail expérimental sur de très nombreux échantillons ayant des origines très diverses (9 428 métagénomes prélevés sur 5 sites corporels sur des individus provenant de 31 pays différents). Le choix de GenBank pour affilier au mieux les espèces connues peut donc se justifier. Mais dans le cadre d'une étude avec un cadre plus restreint, il convient d'utiliser des ressources génomiques de plus grande qualité, voire même de vérifier les séquences récupérées.

L'avancée des techniques de séquençage renforce le risque d'erreur. En effet, auparavant le séquençage d'un génome nécessitait des ressources financières et matérielles particulièrement élevées. Chaque génome était donc affilié et annoté avec la plus grande attention. De nos jours, il est possible

de séquencer un génome, voire un milieu entier, pour un coût bien moindre (Wetterstrand 2019). La répartition des ressources humaines et techniques devrait, de ce fait, aujourd'hui être largement orientée vers l'analyse de la masse de données produites, car elle constitue la partie de l'étude la plus exigeante en terme de temps (Vincent et al. 2017). Cet afflux réduit aussi les possibilités d'affiliations et d'annotations manuelles de ces données et l'utilisation de pipelines permettant l'affiliation et l'annotation automatique est souvent la règle. Cette méthode conduit à la réplication des erreurs. Les erreurs peuvent enfin résulter d'une méconnaissance, par les auteurs de dépôts de génomes dans GenBank, des dernières avancées en matière de classification et de taxonomie.

La rectification de ces erreurs est également un travail peu valorisant et fastidieux à effectuer. En effet, dans les banques de données de l'INSDC (DDBJ, EMBL et NCBI) par exemple, les mises à jour ou corrections d'un dépôt ne sont possibles que par les auteurs initiaux. La correction passe donc par la soumission de nouvelles annotations dans une base de données nommée TPA (Third Party Annotation) et doivent faire l'objet d'une publication avec revue par les pairs (`https://www.ncbi.nlm.nih.gov/books/NBK53704/`). Il n'est pas possible de publier des corrections mineures, à moins de procéder à la correction d'énormes jeux de données.

MACADAM

Disposer de bases de données fiables en terme de classification et d'annotation fonctionnelle, avec une possibilité de mise à jour récurrente, est donc d'un intérêt majeur pour la communauté scientifique. MACADAM a été créé dans le but de disposer d'une base de données regroupant les données fonctionnelles les plus fiables. MACADAM peut servir de base de données d'un outil d'inférence fonctionnelle si cet outil prédit le potentiel fonctionnel à partir d'une information taxonomique comme MACADAMExplore. Parmi les bases de données disponibles au début des travaux de thèse, aucune base de données fonctionnelles ne répondait à ces prérequis : (i) compte tenu de la structure des fichiers de MetaCyc, une interrogation en temps réel de leurs données n'était pas possible. De plus, seules les PGDBs de BioCyc avec une antériorité de deux ans étaient consultables librement. (ii) Le modèle économique proposé par KEGG, basé sur un abonnement, n'offrait pas un accès libre et gratuit aux informations à jour. Seule la version de 2013 peut être récupéré par d'autres voies (la version 2013 est répliquée dans différents pipelines). (iii) PATRIC ne permet pas le téléchargement de ses informations en local. MACADAM a donc été créé pour disposer d'une base de données dont

la mise à jour est facilitée, librement utilisable par la communauté scientifique.

La fonction première de MACADAM est de relier l'information fonctionnelle avec la taxonomie procaryote. La taxonomie à la base de la création de MACADAM provient du NCBI Taxonomy. Ce choix se justifie notamment par le meilleur équilibre entre sa fréquence de mise à jour, sa compatibilité avec les autres bases de données taxonomiques et sa diversité au sens de la représentativité du domaine des procaryotes. A chaque mise à jour de MACADAM, la version actuelle du NCBI Taxonomy est intégrée à MACADAM afin de profiter des nouvelles avancées de la taxonomie et les changements dans la nomenclature. Afin d'éviter au maximum les erreurs d'affiliation taxonomique évoquées dans la partie précédente de la discussion (mauvaise affiliation d'espèces métagénomiques de génomes présents dans RefSeq), les génomes sélectionnés pour inférer les voies métaboliques présentent une qualité d'assemblage « complete genome », et *a priori* une affiliation taxonomique plus fiable que pour des assemblages de moindre qualité (contig, scaffold ou chromosome). Dans notre étude, nous avons ainsi observé que des erreurs d'affiliation étaient liées à l'inclusion dans les bases de génomes sous forme de contig ou scaffold (par exemple dans le genre *Blautia* et dans le genre *Megasphaera*, Partie B.1.5.3, page 96). Cette qualité d'assemblage permet de nous assurer également d'une qualité d'annotation fonctionnelle. En effet, RefSeq labellise certains « complete genome » comme « reference genome » ou « representative genome » profitant, pour les premiers, d'un effort d'annotation manuelle par le NCBI et la communauté de microbiologistes, et d'un contrôle qualité accru pour les seconds (Tatusova et al. 2014). Nous avons utilisé Pathway tool pour générer à partir de ces génomes annotés notre propre base de PGDBs (Pathway/Genome Database). Afin d'enrichir et d'augmenter la qualité des informations fonctionnelles dans MACADAM, MicroCyc, une collection de PGDBs manuellement vérifiées par des experts, a été ajoutée. Nous avons choisi de privilégier cette source d'information pour toutes les espèces redondantes avec notre premier jeu de PGDBs. En effet, en plus de la confiance accordée à cette source compte tenu de l'effort de curation de la communauté de microbiologistes attachée à MicroCyc, nous avons constaté que leurs PGDBs présentaient plus de voies métaboliques que celles que nous avons générées à partir de RefSeq (Figure 3 de l'article MACADAM, page 107).

A cette collection de voies métaboliques sont ajoutées FAPROTAX (Louca, Parfrey et Doebeli 2016) et IJSEM phenotypic database (Barberán et al. 2017), deux bases de données contenant une information fonctionnelle provenant de la littérature. Les métabolites, noms des réactions, enzymes, ainsi que leur nomenclature EC sont également ajouté à MACADAM, permet-

tant de restreindre les recherches. Enfin, nous avons proposé à l'utilisateur de MACADAM le calcul d'un score permettant d'évaluer la probabilité de la présence de chacune des voies dans chacun des organismes : le Pathway Score (PS).

Pour compléter MACADAM et le rendre visible auprès de la communauté des microbiologistes, nous avons développé un script d'interrogation nommé MACADAMExplore (`https://github.com/maloleboulch/MACADAMExplore`) et une interface web associée (`http:\macadam.toulouse.inra.fr`). Outre la simplicité d'interrogation, l'une des originalités de MACADAMExplore est de permettre, en l'absence de données fonctionnelles sur un taxon, de remonter au rang taxonomique supérieur.

MACADAM, bien que générée automatiquement, s'appuie donc sur des sources de données maximisant le degré de qualité de leurs informations et essayant de limiter au maximum les informations erronées et sacrifiant sans aucun doute à l'exhaustivité de la représentation de la diversité procaryote. MACADAM est une base de données librement consultable, gratuite, téléchargeable, avec un code source ouvert, et intégrable dans un outil car se reposant sur des technologies interopérables et libres. MACADAM, se reposant en grande partie sur RefSeq, le NCBI Taxonomy et MetaCyc, ne cessera d'enrichir son contenu et de corriger ses erreurs, notamment taxonomiques, au même rythme que les bases de données dont MACADAM dépend. Nous avons choisi de construire MACADAM sur un modèle relationnel. De ce fait, MACADAM peut inclure de l'information de type méta-liens, permettant la correspondance avec des voies dans d'autres banques de données.

Bien que MACADAM soit conçu pour agréger des sources de données maximisant la qualité des annotations, celle-ci souffre d'une dépendance aux sources de données la composant. En effet, MACADAM est fondée sur les PGDBs et le logiciel Pathway Tools, ce dernier pouvant être amené à changer de méthode, de format de fichier de sortie, ou encore de licence d'utilisation de MetaCyc. MetaCyc ne propose pas de format standard pour les fichiers la composant et ceci peut impliquer des problèmes de prise en charge pour une future version. MACADAM repose également sur l'API (Interface de programmation d'application) de MicroCyc et tout changement de l'API impliquera une refonte de la partie de MACADAM s'occupant de la gestion de ces PGDBs. Afin de limiter l'impact d'une telle refonte, le langage de programmation Python 3 - dernière version de ce langage - et le code source de MACADAM sont disponibles sur GitHub, permettant un support long terme. MACADAM est également sensible aux changements de format, actuellement non standard, ainsi qu'au potentiel changement de licence de la banque de données MetaCyc et de son outil Pathway Tools. MACADAM est enfin soumis aux contraintes de la taxonomie et de l'harmonisation de la no-

menclature. Les *Cyanobacteria* illustrent cette difficulté en lien avec les rangs taxonomiques qui les définissent. En effet, les *Cyanobacteria* ne présentent pas le rang taxonomique « class » dans MACADAM car le NCBI Taxonomy n'en contient pas. La nomenclature des *Cyanobacteria* dépendait auparavant de l'ICN (« International Code of Nomenclature for algae, fungi,and plant ») et les règles de nomenclatures les régissant ne sont pas encore définies formellement entre la communauté des botanistes et des microbiologistes (Oren et Ventura 2017). Cette spécificité a donc été incluse dans MACADAM, qui peut contenir des organismes ne possédant pas l'intégralité des sept rangs taxonomiques majeurs. A l'inverse, certains organismes présentent des rangs optionnels supplémentaires tels que définis dans la partie A.1.6, page 28, et à la figure A.4, page 30. Par exemple, l'organisme *Ruminiclostridium cellobioparum* subsp. *termitidis* appartient à la sous-espèce *termitidis*. Afin de garder un maximum d'homogénéité taxonomique dans MACADAM, ces rangs optionnels ont été écartés lors de la recherche par défaut et de la présentation des résultats. Mais ils sont présents dans la table « taxonomy » de MACADAM et il est possible, en spécifiant une option (« *-nonscientific* »), de rechercher un rang optionnel ainsi que les synonymes, les anciens noms des espèces et les noms vernaculaires. Si cette option est activée, alors le nom scientifique supérieur le plus proche sera utilisé dans les résultats. Par ailleurs, MACADAM dépend aussi des différences de taxonomie entre les différentes bases de données taxonomiques. En effet, bien que le NCBI Taxonomy soit la base de données qui propose le plus de correspondances entre les différentes bases tout en étant associée à une banque de données génomique (GenBank/RefSeq), il existe certaines différences entre les bases de données (Figure A.IX, page 41).

En conclusion, la base de données MACADAM, que nous avons conçue, répond aux objectifs que nous nous sommes fixés au début de notre travail à savoir : libre d'accès, avec un code ouvert, téléchargeable, mettant à disposition des données fonctionnelles volontairement restreintes pour maximiser la fiabilité, tout en limitant les répercussions des erreurs et en faisant bénéficier la communauté des dernières mises à jours taxonomiques et fonctionnelles en répercutant à chaque mise à jour de MACADAM les changements ayant lieu dans le NCBI Taxonomy, les annotations fonctionnelles des génomes RefSeq et de MicroCyc.

Affiliation taxonomique et inférence fonctionnelle

Dans le cadre de ce travail de thèse nous avons réalisé une inférence fonctionnelle d'une communauté procaryote à partir de l'information taxonomique via l'utilisation de MACADAM, et ce, pour deux études différentes : (i) pour l'inférence fonctionnelle de groupes taxonomiques dominants présents dans le microbiote cæcal de lapereaux et (ii) dans l'analyse fonctionnelle des genres *Blautia*, *Ruminococcus* et *Mediterraneibacter*.

L'approche taxonomique que nous avons utilisée dans le cadre de l'étude publiée (Partie B.3.2, page 116, Read et al. 2019) et dans la comparaison du potentiel fonctionnel des genres présents lors de l'étude de l'affiliation taxonomique des espèces métagénomiques (Partie B.3.3, page 120), présente l'avantage d'être précise quant à la requête effectuée. Toutefois, l'information que nous avons obtenue reste limitée. En effet, si les taxons sont bien présents et bien positionnés dans la classification, le faible nombre de « complete genome » sur lesquels s'appuie la construction des PGDBs de MACADAM pénalise fortement l'analyse fonctionnelle différentielle des taxons étudiés. En effet, dans notre étude, comparativement à la base NCBI actuelle, l'inférence n'a été réalisée que sur un pourcentage limité des organismes que constituent ces taxons (Tableau B.13, page 126).

Par ailleurs, l'analyse fonctionnelle à un rang taxonomique élevé, comme la famille par exemple, dépend des organismes qui le composent et toute reclassement de taxons inférieurs impactera le résultat de l'inférence fonctionnelle. Dans notre étude, ce phénomène concerne notamment les espèces *Acetivibrio cellulolyticus*, *[Clostridium] clariflavum*, *[Clostridium] stercorarium*, *Clostridium papyrosolvens*, *Clostridium termitidis*, *Clostridium thermocellum*, *Mageeibacillus indolicus* présents lors de l'étude dans la famille des *Ruminococcaceae*, qui ont depuis été reclassées dans la famille des *Hungateiclostridiaceae* et leurs noms ont évolué en conséquence (Zhang et al. 2018 ; Oren et Garrity 2018a). Notre choix d'inférer l'information fonctionnelle au niveau du rang taxonomique supérieur est ici pris en défaut. La mise à jour de MACADAM prendra en compte ces modifications dès lors qu'elles sont effectives dans le NCBI Taxonomy. L'intérêt d'un système de mise à jour facilité permet de garantir que MACADAM intégrera les évolutions au fil du reclassement des organismes. Une inférence fonctionnelle au plus proche voisin pourrait être une méthode alternative à l'analyse fonctionnelle au rang taxonomique supérieur. Cela nécessite d'identifier les plus proches voisins d'un taxon par une approche phylogénétique. et que ceux-ci possèdent une information fonctionnelle qui leur est associée. Les intérêts

respectifs de ces deux approches restent à étudier.

L'analyse du potentiel fonctionnel des genres *Blautia*, *Ruminococcus* et *Mediterraneibacter* (Figure B.XI, page 125) ne permet pas de regrouper les espèces selon leur phylogénie, mais plutôt en fonction du nombre de voies métaboliques annotées et plus particulièrement en fonction de l'origine de leur génome ou de l'origine de la PGDB (RefSeq vs MicroCyc). Ce résultat met en exergue la problématique des voies présentes dans l'organisme mais absentes dans les résultats (on parle de faux négatifs), et des problèmes générés par les différences de qualité d'annotation entre organismes peu connus et organismes de référence.

Lecture métagénomique et inférence fonctionnelle

Dans le cadre de notre travail de thèse nous avons également confronté deux approches méthodologiques pour réaliser une inférence fonctionnelle (Partie B.4, page 128) : (i) à partir d'une information taxonomique via l'utilisation de MACADAM et (ii) à partir de lectures métagénomiques via l'utilisation de l'outil HUMAnN2. Pour réaliser cette comparaison, nous nous sommes basés sur une communauté bactérienne artificielle de 11 espèces bactériennes présentées dans le tableau B.14, page 129. Une méthodologie permettant de comparer des voies identifiées dans les résultats produits à partir de MACADAM et de HUMAnN2 a été mise en place.

Le pipeline HUMAnN2 tend à retrouver les 11 espèces bactériennes présentes dans le milieu grâce à sa base de données de pangénomes annotés : ChocoPhlAn. Par ailleurs, une espèce (*Paeniclostridium sordellii*) est présente sous son ancien nom (*Clostridium sordellii*) dans cette base de données, contrairement à MACADAM. Bien que HUMAnN2 affiche le mauvais nom, les résultats n'en dépendent pas pour inférer les voies métaboliques présentes dans le génome.

Le nombre de voies identifiées dans les PGDBs - à partir du nom des espèces par MACADAM et à partir des lectures par HUMAnN2 - a été analysé B.XV, page 138. MACADAM présente plus de voies métaboliques que HUMAnN2. Cette différence peut s'expliquer par la moindre qualité d'assemblage des génomes composant la base de données ChocoPhlAn, comparativement à MACADAM. Une autre explication concerne la taille des lectures sur lesquelles se base HUMAnN2 pour inférer les voies métaboliques, ainsi que la couverture et/ou la profondeur de séquençage. MACADAM se reposant sur une inférence fonctionnelle à partir de noms taxonomiques, elle n'est pas limitée par ces biais de technologies de séquençage.

Toutefois, les limites de cette étude comparative sont multiples : (i) la communauté bactérienne artificielle est composée d'un nombre réduit d'es-

pèces, et celles-ci sont présentes dans MACADAM. Dans le cas contraire, MACADAM serait remonté au rang supérieur. En l'absence de génome de référence dans ChocoPhlan, HUMAnN2 permet une inférence fonctionnelle en ayant recours à son outil de recherche par des séquences nucléiques traduites, avec comme seule limite les banques de données de protéines connues. (ii) Nous avons fait le choix de n'appliquer aucun filtre de score aux résultats des deux méthodes. Le pathway score (PS) et le pathway frequency score (PFS) de MACADAM, ainsi que le pathway coverage et le pathway abundance de HUMAnN2 n'ont pas été pris en compte. Des 714 voies de MACADAM et des 286 de HUMAnN2, de nombreuses pourraient, de ce fait, être de faux positifs. Il conviendrait de trouver une correspondance entre le PS et le pathway coverage afin d'éliminer les faux positifs des voies métaboliques présentes dans les résultats.

Limites de l'affiliation taxonomique et de l'inférence fonctionnelle

L'ensemble des travaux que nous avons conduits souligne l'importance de la qualité d'assemblage et d'annotation des génomes produits pour réaliser des études d'inférence métabolique ou d'assignation taxonomique. En effet, la qualité des génomes présents dans les bases dépend de la qualité et/ou fiabilité des assemblages et des annotations. Ces informations se répercutent sur le positionnement des organismes lors de la reconstruction d'arbres phylogénétiques à partir de protéines marqueurs. Par ailleurs, nous avons observé qu'une qualité d'annotation insuffisante pouvait conduire à la production de faux négatifs, c'est à dire des voies notées absentes, car les enzymes qui les composent sont mal annotées dans les génomes (Schnoes et al. 2009). A l'inverse, un intérêt accru pour un organisme conduit à une meilleure annotation. Cette haute qualité d'annotation peut alors biaiser l'importance fonctionnelle de cet organisme au sein de sa communauté.

L'importance d'utiliser des bases de données répercutant les dernières mises à jours est également cruciale. C'est le cas pour les bases de données taxonomiques, afin de capter au mieux la diversité procaryotique ou les avancées de la systématique. C'est également le cas pour les bases de données métaboliques : l'obsolescence de l'information conduit à des erreurs et/ou à la sous-estimation du potentiel fonctionnel (Wadi et al. 2016). Wadi et al. (2016) conseillent un rythme de mise à jour de six mois pour une base de données fonctionnelles, afin de suivre les avancées scientifiques. MACADAM étant basé sur RefSeq et sur le NCBI Taxonomy, nous comptons suivre ce rythme comme indiqué dans la partie « Management of the MACADAM

database update » de l'article MACADAM, page 109. Par sa nature open source et son interopérabilité grâce à l'utilisation de Python 3 et SQLite, le support de MACADAM est facilité sur le long terme si des ressources humaines sont disponibles afin d'effectuer la mise à jour régulière. Ce support dépend également du support apporté par leurs auteurs aux bases de données fonctionnelles (MetaCyc, MicroCyc) et ainsi qu'a l'outil Pathway Tools.

Enfin, il est important de préciser que quelle que soit la méthode d'inférence utilisée pour les travaux effectués dans ce travail de thèse (inférence à partir de la taxonomie ou à partir de lectures), il est seulement possible de parler de potentiel fonctionnel (Partie A.3.4, page 65), c'est-à-dire quelles sont les fonctions et/ou les voies métaboliques que l'organisme ou la communauté procaryotique est en mesure de réaliser. Ne possédant aucune donnée d'expression de gènes, nous sommes dans l'impossibilité d'être certains qu'une voie est activée par un organisme plutôt que par un autre, par deux organismes en même temps, ou encore que cette voie soit présente mais non activée.

Conclusions

Ce travail de thèse avait pour objectif de répondre à trois problématiques.

La première problématique concerne la nécessité de disposer d'une base de données fonctionnelles de haute qualité, en accès libre et interopérable afin mieux comprendre les fonctions des organismes procaryotiques. Au démarrage de notre travail de thèse, cette base de données n'était pas disponible, c'est pourquoi nous avons créé la base de données MACADAM. La deuxième problématique que nous avons étudiée concerne la comparaison de résultats d'inférences de voie métabolique à partir d'informations taxonomiques ou de lectures métagénomiques. Cette comparaison présente des résultats encourageants pour la recherche par noms taxonomiques mais celle-ci mériterait d'être étendue et précisée. La dernière concerne la correction d'assignation taxonomique d'espèces métagénomiques en utilisant une approche par reconstruction d'un arbre phylogénétique d'une part et en utilisant un indice global de parenté génomique d'autre part. Ce travail a permis de montrer que l'affiliation d'espèces métagénomiques nécessite un travail de curation principalement lié à des problèmes de nomenclature et de qualité d'annotation taxonomique dans les bases de données. La communauté scientifique doit se doter de moyens pour respecter la nomenclature, curer les bases de données et acquérir de nouvelles connaissances sur les espèces procaryotiques. Cela inclus des efforts relatifs à la culture d'espèce aujourd'hui encore incultivable, de séquençage de leur génome et de la qualité de leur annotation (faux négatifs et faux positifs variabilité dans la profondeur des annotations). Par ce travail, outre le repositionnement des neuf groupes d'espèces métagénomiques, nous contribuons modestement au reclassement de génomes dans *Blautia obeum*, *Megasphaera stantonii* et nous confirmons que *Ruminococcus gauvreauii* devrait être déplacé dans le genre *Blautia*.

Perspectives

MACADAM a été conçu pour être intégrer dans des outils notamment d'inférence fonctionnelle à partir de gènes marqueurs. De part son interroga-

tion en SQLite et par l'utilisation du NCBI Taxonomy celle-ci est intégrable à la suite d'outils d'inférence fonctionnelle à partir de gènes marqueurs si ceux-ci prédisent le potentiel fonctionnel à partir d'information taxonomique. Bien que MACADAM soit facile à requêter, la forme des résultats (un fichier tabulé) n'est pas assez riche afin de rallier un maximum d'utilisateurs. MACADAM en plus de ces mises à jour régulières pourrait être étendue à davantage de méta données. MACADAM gagnerait à intégrer des liens avec une banque de données protéiques notamment UniProt. La communauté scientifique utilisant également les voies métaboliques de KEGG ainsi que les KEGG Orthology, celles-ci seraient intéressante à associées aux voies métaboliques présentes dans MACADAM.

Par ailleurs, une refonte plus profonde de MACADAM serait intéressante afin d'apporter plus de diversité dans les voies métaboliques ainsi qu'une plus grande précision dans les scores de complétude des voies métaboliques. Cette refonte permettrait également de ne pas être dépendant du modèle économique de MetaCyc et de Pathway Tools, outil très utile mais pouvant être obscur dans son fonctionnement. Pour cela, lors du téléchargement des génomes RefSeq, en parallèle de la construction des PGDBs par Pathway Tools, il serait intéressant d'utiliser entre autres Prokka (Seemann 2014), eggNOG-mapper (Huerta-Cepas et al. 2017) ou COBRApy (Ebrahim et al. 2013) afin de compléter au maximum les annotations des génomes RefSeq et de relier ces annotations aux identifiants Uniprot qui seraient présents dans MACADAM. L'ensemble de ces annotations serviraient de source d'inférence de voies métaboliques provenant de MetaCyc grâce a MinPath (Ye et Doak 2009). Par la suite, il serait possible de calculer un score de couverture des voies métaboliques comme dans HUMAnN2 et de comparer les voies métaboliques de cette méthode avec les voies métaboliques présentes dans les PGDBs.

Il est possible d'utiliser MACADAMExplore afin d'interroger MACADAM sur plusieurs organismes en une seule demande, il serait donc intéressant que MACADAMExplore puisse se baser sur une table d'abondance issue d'un outil d'affiliation taxonomique et d'inclure un post-traitement permettant de prendre en compte la participation de chacun des organismes et de mettre en avant les relations de syntrophie au sein de la communauté bactérienne.

La taxonomie devrait également être un sujet traité avec davantage d'intérêt lors de séquençage de génomes et/ou de métagénomes. La communauté scientifique devrait en outre être plus au fait de la nomenclature et des méthodes de systématique de son domaine du vivant et s'astreindre à un meilleur respect des règles de nomenclature. Les efforts fait pour corriger les données présentes dans les bases de données, cultiver les organismes types, les séquen-

cer et les annoter devraient également être sujet à un meilleur financement et de meilleures possibilités de valorisation. Enfin bien que la métagénomique soit un outil qui nous permette de mieux comprendre les milieux qui nous entoure et leurs compositions, il ne faut pas se reposer seulement sur cette méthode. Un effort de culture des organismes réputés incultivables et le séquençage de ces colonies permettraient d'accroître les connaissances et la diversité des organismes procaryotes dans les banques de données.

Bibliographie

Abubucker, S., N. Segata, J. Goll, A. M. Schubert, J. Izard, B. L. Cantarel, B. Rodriguez-Mueller, J. Zucker, M. Thiagarajan, B. Henrissat, O. White, S. T. Kelley, B. Methé, P. D. Schloss, D. Gevers, M. Mitreva et C. Huttenhower (13 juin 2012). « Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome ». In : *PLoS Computational Biology* 8.6. issn : 1553-734X. doi : 10.1371/journal.pcbi.1002358.

Achtman, M. et M. Wagner (juin 2008). « Microbial diversity and the genetic nature of microbial species ». In : *Nature Reviews Microbiology* 6.6, p. 431-440. issn : 1740-1534. doi : 10.1038/nrmicro1872.

Acinas, S. G., L. A. Marcelino, V. Klepac-Ceraj et M. F. Polz (mai 2004). « Divergence and Redundancy of 16S rRNA Sequences in Genomes with Multiple rrn Operons ». In : *Journal of Bacteriology* 186.9. Citation Key Alias : acinas_divergence_2004-1, p. 2629-2635. issn : 0021-9193. doi : 10.1128/JB.186.9.2629-2635.2004.

Altman, T., M. Travers, A. Kothari, R. Caspi et P. D. Karp (27 mar. 2013). « A systematic comparison of the MetaCyc and KEGG pathway databases ». In : *BMC Bioinformatics* 14, p. 112. issn : 1471-2105. doi : 10.1186/1471-2105-14-112.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers et D. J. Lipman (5 oct. 1990). « Basic local alignment search tool ». In : *Journal of molecular biology* 215.3, p. 403-410. issn : 0022-2836. doi : 10.1016/S0022-2836(05)80360-2.

Amann, R., W. Ludwig et K. H. Schleifer (1er mar. 1995). « Phylogenetic identification and in situ detection of individual microbial cells without cultivation. » In : *Microbiological Reviews* 59.1, p. 143-169. issn : 1092-2172, 1098-5557.

Amir, A., D. McDonald, J. A. Navas-Molina, E. Kopylova, J. T. Morton, Z. Z. Xu, E. P. Kightley, L. R. Thompson, E. R. Hyde,

A. Gonzalez et R. Knight (21 avr. 2017). « Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns ». In : *mSystems* 2.2, e00191-16. issn : 2379-5077. doi : 10.1128/mSystems.00191-16.

Andrews, S. (2010). *FastQC : a quality control tool for high throughput sequence data.* url : http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Arahal, D. R. (1er jan. 2014). « Chapter 6 - Whole-Genome Analyses : Average Nucleotide Identity ». In : *Methods in Microbiology*. Sous la dir. de M. Goodfellow, I. Sutcliffe et J. Chun. T. 41. New Approaches to Prokaryotic Systematics. Academic Press, p. 103-122. doi : 10.1016/bs.mim.2014.07.002.

Arenas, M. (26 oct. 2015). « Trends in substitution models of molecular evolution ». In : *Frontiers in Genetics* 6. issn : 1664-8021. doi : 10.3389/fgene.2015.00319.

Asshauer, K. P. et P. Meinicke (2013). « On the estimation of metabolic profiles in metagenomics ». In : *German conference on bioinformatics 2013*. Sous la dir. de T. Beissbarth, M. Kollmar, A. Leha, B. Morgenstern, A.-K. Schultz, S. Waack et E. Wingender. T. 34. OpenAccess series in informatics (OASIcs). Dagstuhl, Germany : Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, p. 1-13. isbn : 978-3-939897-59-0. doi : 10.4230/OASIcs.GCB.2013.1.

Asshauer, K. P., B. Wemheuer, R. Daniel et P. Meinicke (1er sept. 2015). « Tax4Fun : predicting functional profiles from metagenomic 16S rRNA data ». In : *Bioinformatics* 31.17, p. 2882-2884. issn : 1367-4803, 1460-2059. doi : 10.1093/bioinformatics/btv287.

Aubert, D. (juil. 2016). « Doit-on parler de « nomenclature binomiale » ou bien de « nomenclature binominale » ». In : *La banque des mots* 91, p. 7-14.

Auch, A. F., M. von Jan, H.-P. Klenk et M. Göker (28 jan. 2010). « Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison ». In : *Standards in Genomic Sciences* 2.1, p. 117-134. issn : 1944-3277. doi : 10.4056/sigs.531120.

Ayling, M., M. D. Clark et R. M. Leggett (2019). « New approaches for metagenome assembly with short reads ». In : *Briefings in bioinformatics*.

Balvočiūtė, M. et D. H. Huson (14 mar. 2017). « SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare ? » In : *BMC*

Genomics 18.2, p. 114. ISSN : 1471-2164. DOI : 10.1186/s12864-017-3501-4.

Barbera, P., A. M. Kozlov, L. Czech, B. Morel, D. Darriba, T. Flouri et A. Stamatakis (1er mar. 2019). « EPA-ng : Massively Parallel Evolutionary Placement of Genetic Sequences ». In : *Systematic Biology* 68.2, p. 365-369. ISSN : 1063-5157. DOI : 10.1093/sysbio/syy054.

Barberan, A. (6 déc. 2016). *International Journal of Systematic and Evolutionary Microbiology (IJSEM) phenotypic database*. DOI : 10.6084/m9.figshare.4272392.v3.

Barberán, A., H. C. Velazquez, S. Jones et N. Fierer (30 août 2017). « Hiding in Plain Sight : Mining Bacterial Species Records for Phenotypic Trait Information ». In : *mSphere* 2.4, e00237-17. ISSN : 2379-5042. DOI : 10.1128/mSphere.00237-17.

Bauer, H., H. Giebl, R. Hitzenberger, A. Kasper-Giebl, G. Reischl, F. Zibuschka et H. Puxbaum (2003). « Airborne bacteria as cloud condensation nuclei ». In : *Journal of Geophysical Research : Atmospheres* 108 (D21). ISSN : 2156-2202. DOI : 10.1029/2003JD003545.

Beaz-Hidalgo, R., M. J. Hossain, M. R. Liles et M.-J. Figueras (21 jan. 2015). « Strategies to Avoid Wrongly Labelled Genomes Using as Example the Detected Wrong Taxonomic Affiliation for Aeromonas Genomes in the GenBank Database ». In : *PLOS ONE* 10.1, e0115813. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0115813.

Beiko, R. G., T. J. Harlow et M. A. Ragan (4 oct. 2005). « Highways of gene sharing in prokaryotes ». In : *Proceedings of the National Academy of Sciences* 102.40, p. 14332-14337. ISSN : 0027-8424, 1091-6490. DOI : 10.1073/pnas.0504068102.

Bengtsson-Palme, J. (1er jan. 2018). « Chapter 3 - Strategies for Taxonomic and Functional Annotation of Metagenomes ». In : *Metagenomics*. Sous la dir. de M. Nagarajan. Academic Press, p. 55-79. ISBN : 978-0-08-102268-9. DOI : 10.1016/B978-0-08-102268-9.00003-3.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell et E. W. Sayers (jan. 2009). « GenBank ». In : *Nucleic Acids Research* 37 (Database issue), p. D26-31. ISSN : 1362-4962. DOI : 10.1093/nar/gkn723.

Bérdy, J. (jan. 2005). « Bioactive Microbial Metabolites ». In : *The Journal of Antibiotics* 58.1, p. 1-26. ISSN : 1881-1469. DOI : 10.1038/ja.2005.1.

Bergey, D. H. (1923). *Bergey's manual of determinative bacteriology*. OCLC : 61582384. Baltimore : Williams & Wilkins Co.

Bergey, D. H., R. E. Buchanan, N. E. Gibbons et American Society for Microbiology (1974). *Bergey's manual of determinative bacteriology*. OCLC : 754547. Baltimore : Williams & Wilkins Co. isbn : 978-0-683-01117-3.

Bolyen, E. et al. (août 2019). « Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2 ». In : *Nature Biotechnology* 37.8, p. 852-857. issn : 1546-1696. doi : 10.1038/s41587-019-0209-9.

Bourne, P. E., J. R. Lorsch et E. D. Green (5 nov. 2015). « Perspective : Sustaining the big-data ecosystem ». In : *Nature* 527.7576, S16-S17. issn : 0028-0836. doi : 10.1038/527S16a.

Bowman, J. S. et H. W. Ducklow (2015). « Microbial Communities Can Be Described by Metabolic Structure : A General Framework and Application to a Seasonally Variable, Depth-Stratified Microbial Community from the Coastal West Antarctic Peninsula ». In : *PLOS ONE* 10.8, e0135868. issn : 1932-6203. doi : 10.1371/journal.pone.0135868.

Breitwieser, F. P., J. Lu et S. L. Salzberg (23 sept. 2017). « A review of methods and databases for metagenomic classification and assembly ». In : *Briefings in Bioinformatics*. issn : 1477-4054. doi : 10.1093/bib/bbx120.

Brenner, D. J., G. R. Fanning, A. V. Rake et K. E. Johnson (1er jan. 1969). « Batch procedure for thermal elution of DNA from hydroxyapatite ». In : *Analytical Biochemistry* 28, p. 447-459. issn : 0003-2697. doi : 10.1016/0003-2697(69)90199-7.

Brosius, J., M. L. Palmer, P. J. Kennedy et H. F. Noller (oct. 1978). « Complete nucleotide sequence of a 16S ribosomal RNA gene from Escherichia coli. » In : *Proceedings of the National Academy of Sciences of the United States of America* 75.10, p. 4801-4805. issn : 0027-8424.

Brown, D. R., R. F. Whitcomb et J. M. Bradbury (2007). « Revised minimal standards for description of new species of the class Mollicutes (division Tenericutes) ». In : *International Journal of Systematic and Evolutionary Microbiology,* 57.11, p. 2703-2719. issn : 1466-5026, doi : 10.1099/ijs.0.64722-0.

Browne, H. P., S. C. Forster, B. O. Anonye, N. Kumar, B. A. Neville, M. D. Stares, D. Goulding et T. D. Lawley (mai 2016). « Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation. » In : *Nature* 533.7604, p. 543-546. issn : 0028-0836. doi : 10.1038/nature17645.

Bruto, M. (1er sept. 2010). « Étude de la plasticité génomique chez Streptomyces ambofaciens : assemblage et analyse comparative du génome des souches ATCC23877 et DSM40697 ». Thèse de doct.

Buchanan, R. E., R. T. St. John-Brooks et International Congress for Microbiology (1947). *Proposed bacteriological code of nomenclature developed from proposals approved by the International Committee on Bacteriological Nomenclature at the meeting of the Third International Congress for Microbiology*. Ames, Ia. : Iowa State College Press. 7-61 p.

Buchfink, B., C. Xie et D. H. Huson (jan. 2015). « Fast and sensitive protein alignment using DIAMOND ». In : *Nature Methods* 12.1, p. 59-60. issn : 1548-7105. doi : 10.1038/nmeth.3176.

Bui, T. P. N., J. Ritari, S. Boeren, P. de Waard, C. M. Plugge et W. M. de Vos (1er déc. 2015). « Production of butyrate from lysine and the Amadori product fructoselysine by a human gut commensal ». In : *Nature Communications* 6. issn : 2041-1723. doi : 10.1038/ncomms10062.

Bukin, Y. S., Y. P. Galachyants, I. V. Morozov, S. V. Bukin, A. S. Zakharenko et T. I. Zemskaya (5 fév. 2019). « The effect of 16S rRNA region choice on bacterial community metabarcoding results ». In : *Scientific Data* 6, p. 190007. issn : 2052-4463. doi : 10.1038/sdata.2019.7.

Cabrera, M. Á. et J. M. Blamey (5 oct. 2018). « Biotechnological applications of archaeal enzymes from extreme environments ». In : *Biological Research* 51.1, p. 37. issn : 0717-6287. doi : 10.1186/s40659-018-0186-3.

Callahan, B. J., P. J. McMurdie et S. P. Holmes (déc. 2017). « Exact sequence variants should replace operational taxonomic units in marker-gene data analysis ». In : *The ISME Journal* 11.12, p. 2639-2643. issn : 1751-7370. doi : 10.1038/ismej.2017.119.

Callahan, B. J., P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson et S. P. Holmes (juil. 2016). « DADA2 : High-resolution sample inference from Illumina amplicon data ». In : *Nature Methods* 13.7, p. 581-583. issn : 1548-7091. doi : 10.1038/nmeth.3869.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer et T. L. Madden (15 déc. 2009). « BLAST+ : architecture and applications ». In : *BMC Bioinformatics* 10.1, p. 421. issn : 1471-2105. doi : 10.1186/1471-2105-10-421.

Candolle, A. P. d. (1813). *Théorie élémentaire de la botanique ; ou, Exposition des prinicpes de la classification naturelle et de l'art de décrire et d'étudier les végétaux.*

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld et R. Knight (mai 2010). « QIIME allows analysis of high-throughput community sequencing data ». In : *Nature methods* 7.5, p. 335-336. issn : 1548-7091. doi : 10.1038/nmeth.f.303.

Case, R. J., Y. Boucher, I. Dahllof, C. Holmstrom, W. F. Doolittle et S. Kjelleberg (jan. 2007). « Use of 16S rRNA and rpoB Genes as Molecular Markers for Microbial Ecology Studies ». In : *Applied and Environmental Microbiology* 73.1, p. 278-288. issn : 0099-2240. doi : 10.1128/AEM.01177-06.

Caspi, R., T. Altman, R. Billington, K. Dreher, H. Foerster, C. A. Fulcher, T. A. Holland, I. M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, D. Weerasinghe, P. Zhang et P. D. Karp (1er jan. 2014). « The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases ». In : *Nucleic Acids Research* 42 (D1), p. D459-D471. issn : 0305-1048, 1362-4962. doi : 10.1093/nar/gkt1103.

Caspi, R., R. Billington, L. Ferrer, H. Foerster, C. A. Fulcher, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver et P. D. Karp (4 jan. 2016). « The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases ». In : *Nucleic Acids Research* 44 (D1), p. D471-D480. issn : 0305-1048. doi : 10.1093/nar/gkv1164.

Caspi, R., R. Billington, C. A. Fulcher, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, P. E. Midford, Q. Ong, W. K. Ong, S. Paley, P. Subhraveti et P. D. Karp (4 jan. 2018). « The MetaCyc database of metabolic pathways and enzymes ». In : *Nucleic Acids Research* 46 (D1), p. D633-D639. issn : 0305-1048. doi : 10.1093/nar/gkx935.

Cavalier-Smith, T. (29 juin 2006). « Cell evolution and Earth history : stasis and revolution ». In : *Philosophical Transactions of the Royal Society B : Biological Sciences* 361.1470, p. 969-1006. issn : 0962-8436. doi : 10.1098/rstb.2006.1842.

Cavicchioli, R. (jan. 2011). « Archaea — timeline of the third domain ». In : *Nature Reviews Microbiology* 9.1, p. 51-61. issn : 1740-1534. doi : 10.1038/nrmicro2482.

Chaban, B., S. Y. Ng et K. F. Jarrell (1er fév. 2006). « Archaeal habitats — from the extreme to the ordinary ». In : *Canadian Journal of Microbiology* 52.2, p. 73-116. issn : 0008-4166. doi : 10.1139/w05-147.

Chakravorty, S., D. Helb, M. Burday, N. Connell et D. Alland (mai 2007). « A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria ». In : *Journal of Microbiological Methods* 69.2, p. 330-339. issn : 0167-7012. doi : 10.1016/j.mimet.2007.02.005.

Chen, I.-M. A., K. Chu, K. Palaniappan, M. Pillay, A. Ratner, J. Huang, M. Huntemann, N. Varghese, J. R. White, R. Seshadri, T. Smirnova, E. Kirton, S. P. Jungbluth, T. Woyke, E. A. Eloe-Fadrosh, N. N. Ivanova et N. C. Kyrpides (8 jan. 2019). « IMG/M v.5.0 : an integrated data management and comparative analysis system for microbial genomes and microbiomes ». In : *Nucleic Acids Research* 47 (Database issue), p. D666. doi : 10.1093/nar/gky901.

Chun, J., J.-H. Lee, Y. Jung, M. Kim, S. Kim, B. K. Kim et Y.-W. Lim (2007). « EzTaxon : a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences ». In : *International Journal of Systematic and Evolutionary Microbiology* 57.10, p. 2259-2261. doi : 10.1099/ijs.0.64915-0.

Chun, J., A. Oren, A. Ventosa, H. Christensen, D. R. Arahal, M. S. da Costa, A. P. Rooney, H. Yi, X.-W. Xu, S. De Meyer et M. E. Trujillo (2018). « Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes ». In : *International Journal of Systematic and Evolutionary Microbiology* 68.1, p. 461-466. doi : 10.1099/ijsem.0.002516.

Chun, J. et F. A. Rainey (2014). « Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea ». In : *International Journal of Systematic and Evolutionary Microbiology* 64.2, p. 316-324. doi : 10.1099/ijs.0.054171-0.

Claudel-Renard, C., C. Chevalet, T. Faraut et D. Kahn (15 nov. 2003). « Enzyme-specific profiles for genome annotation : PRIAM ». In : *Nucleic Acids Research* 31.22, p. 6633-6639. issn : 1362-4962.

Cohan, F. M. (2002). « What are bacterial species ? » In : *Annual Review of Microbiology* 56, p. 457-487. issn : 0066-4227. doi : 10.1146/annurev.micro.56.012302.160634.

Cole, J. R., Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske et J. M. Tiedje (1er jan. 2014). « Ribosomal Database Project : data and tools for high throughput rRNA analysis ». In : *Nucleic Acids Research* 42 (Database issue), p. D633-D642. issn : 0305-1048. doi : 10.1093/nar/gkt1244.

Collins, M. D. et D. Jones (juin 1981). « Distribution of isoprenoid quinone structural types in bacteria and their taxonomic implication. » In : *Microbiological Reviews* 45.2, p. 316-354. issn : 0146-0749.

Colwell, R. R. (oct. 1970). « Polyphasic Taxonomy of the Genus Vibrio : Numerical Taxonomy of Vibrio cholerae, Vibrio parahaemolyticus, and Related Vibrio Species ». In : *Journal of Bacteriology* 104.1, p. 410-433. issn : 0021-9193.

Consortium, T. U. (8 jan. 2019). « UniProt : a worldwide hub of protein knowledge ». In : *Nucleic Acids Research* 47 (D1), p. D506-D515. issn : 0305-1048. doi : 10.1093/nar/gky1049.

Cornish-Bowden, A. et M. L. Cárdenas (7 déc. 2017). « Life before LUCA ». In : *Journal of Theoretical Biology*. The origin of mitosing cells : 50th anniversary of a classic paper by Lynn Sagan (Margulis) 434, p. 68-74. issn : 0022-5193. doi : 10.1016/j.jtbi.2017.05.023.

Costello, E. K., C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon et R. Knight (18 déc. 2009). « Bacterial community variation in human body habitats across space and time ». In : *Science (New York, N.Y.)* 326.5960, p. 1694-1697. issn : 1095-9203. doi : 10.1126/science.1177486.

Crosa, J. H., D. J. Brenner et S. Falkow (sept. 1973). « Use of a single-strand specific nuclease for analysis of bacterial and plasmid deoxyribonucleic acid homo- and heteroduplexes ». In : *Journal of Bacteriology* 115.3, p. 904-911. issn : 0021-9193.

Czepiel, J., M. Dróżdż, H. Pituch, E. J. Kuijper, W. Perucki, A. Mielimonka, S. Goldman, D. Wultańska, A. Garlicki et G. Biesiada (2019). « Clostridium difficile infection : review ». In : *European Journal*

of Clinical Microbiology & Infectious Diseases 38.7, p. 1211-1221. ISSN : 0934-9723. DOI : 10.1007/s10096-019-03539-6.

Dacks, J. B., M. C. Field, R. Buick, L. Eme, S. Gribaldo, A. J. Roger, C. Brochier-Armanet et D. P. Devos (15 oct. 2016). « The changing view of eukaryogenesis – fossils, cells, lineages and how they all come together ». In : *J Cell Sci* 129.20, p. 3695-3703. ISSN : 0021-9533, 1477-9137. DOI : 10.1242/jcs.178566.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, Or, The Preservation of Favoured Races in the Struggle for Life*. J. Murray. 548 p.

De Vos, P. (1er jan. 2011). « 17 - Multilocus Sequence Determination and Analysis ». In : *Methods in Microbiology*. Sous la dir. de F. Rainey et A. Oren. T. 38. Taxonomy of Prokaryotes. Academic Press, p. 385-407. DOI : 10.1016/B978-0-12-387730-7.00017-6.

Deloger, M., M. El Karoui et M.-A. Petit (jan. 2009). « A Genomic Distance Based on MUM Indicates Discontinuity between Most Bacterial Species and Genera ». In : *Journal of Bacteriology* 191.1, p. 91-99. ISSN : 0021-9193. DOI : 10.1128/JB.01202-08.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu et G. L. Andersen (1er juil. 2006). « Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB ». In : *Applied and Environmental Microbiology* 72.7, p. 5069-5072. ISSN : 0099-2240, 1098-5336. DOI : 10.1128/AEM.03006-05.

Douglas, G. M., R. G. Beiko et M. G. I. Langille (2018). « Predicting the Functional Potential of the Microbiome from Marker Genes Using PICRUSt ». In : *Microbiome Analysis : Methods and Protocols*. Sous la dir. de R. G. Beiko, W. Hsiao et J. Parkinson. Methods in Molecular Biology. New York, NY : Springer New York, p. 169-177. ISBN : 978-1-4939-8728-3. DOI : 10.1007/978-1-4939-8728-3_11.

Douglas, G. M., V. J. Maffei, J. Zaneveld, S. N. Yurgel, J. R. Brown, C. M. Taylor, C. Huttenhower et M. G. I. Langille (15 juin 2019). « PICRUSt2 : An improved and extensible approach for metagenome inference ». In : *bioRxiv*, p. 672295. DOI : 10.1101/672295.

Durand, G. A., T. Pham, S. Ndongo, S. I. Traore, G. Dubourg, J.-C. Lagier, C. Michelle, N. Armstrong, P.-E. Fournier, D. Raoult et M. Million (1er fév. 2017). « Blautia massiliensis sp. nov., isolated from a fresh human fecal sample and emended description of the genus

Blautia ». In : *Anaerobe* 43, p. 47-55. ISSN : 1075-9964. DOI : 10.1016/j.anaerobe.2016.12.001.

Dykhuizen, D. E. et L. Green (nov. 1991). « Recombination in Escherichia coli and the definition of biological species. » In : *Journal of Bacteriology* 173.22, p. 7257-7268. ISSN : 0021-9193.

Dykhuizen, D. E. et G. Baranton (1er juil. 2001). « The implications of a low rate of horizontal transfer in Borrelia ». In : *Trends in Microbiology* 9.7, p. 344-350. ISSN : 0966-842X. DOI : 10.1016/S0966-842X(01)02066-2.

Ebrahim, A., J. A. Lerman, B. O. Palsson et D. R. Hyduke (8 août 2013). « COBRApy : COnstraints-Based Reconstruction and Analysis for Python ». In : *BMC Systems Biology* 7.1, p. 74. ISSN : 1752-0509. DOI : 10.1186/1752-0509-7-74.

Edgar, R. C. (1er oct. 2010). « Search and clustering orders of magnitude faster than BLAST ». In : *Bioinformatics* 26.19, p. 2460-2461. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btq461.

Edgar, R. C. (oct. 2013). « UPARSE : highly accurate OTU sequences from microbial amplicon reads ». In : *Nature Methods* 10.10, p. 996-998. ISSN : 1548-7105. DOI : 10.1038/nmeth.2604.

Edgar, R. C. (12 juin 2018a). « Taxonomy annotation and guide tree errors in 16S rRNA databases ». In : *PeerJ* 6. ISSN : 2167-8359. DOI : 10.7717/peerj.5030.

Edgar, R. C. (15 juil. 2018b). « Updating the 97% identity threshold for 16S ribosomal RNA OTUs ». In : *Bioinformatics* 34.14, p. 2371-2375. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/bty113.

Emwas, A.-H. M. (2015). « The Strengths and Weaknesses of NMR Spectroscopy and Mass Spectrometry with Particular Focus on Metabolomics Research ». In : *Metabonomics : Methods and Protocols*. Sous la dir. de J. T. Bjerrum. Methods in Molecular Biology. New York, NY : Springer New York, p. 161-193. ISBN : 978-1-4939-2377-9. DOI : 10.1007/978-1-4939-2377-9_13.

Ereshefsky, M. (1992). *The Units of Evolution : Essays on the Nature of Species*. Google-Books-ID : uanzAYFCeokC. MIT Press. 432 p. ISBN : 978-0-262-05044-9.

Escudié, F., L. Auer, M. Bernard, M. Mariadassou, L. Cauquil, K. Vidal, S. Maman, G. Hernandez-Raquet, S. Combes et G. Pascal (15 avr. 2018). « FROGS : Find, Rapidly, OTUs with Galaxy Solution ».

In : *Bioinformatics* 34.8, p. 1287-1294. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btx791.

Euzéby, J. P. (1997). « List of Bacterial Names with Standing in Nomenclature : a Folder Available on the Internet ». In : *International Journal of Systematic and Evolutionary Microbiology,* 47.2, p. 590-592. ISSN : 1466-5026, DOI : 10.1099/00207713-47-2-590.

Euzéby, J. (2010). « List of new names and new combinations previously effectively, but not validly, published ». In : *International Journal of Systematic and Evolutionary Microbiology* 60.3, p. 469-472. DOI : 10.1099/ijs.0.022855-0.

Fabregat, A., S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, M. Milacic, C. D. Roca, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, G. Viteri, J. Weiser, G. Wu, L. Stein, H. Hermjakob et P. D'Eustachio (4 jan. 2018). « The Reactome Pathway Knowledgebase ». In : *Nucleic Acids Research* 46 (Database issue), p. D649-D655. ISSN : 0305-1048. DOI : 10.1093/nar/gkx1132.

Falentin, H., L. Auer, M. Mariadassou, G. Pascal, O. Rué, E. DUGAT-BONY, C. Delbes, A. Nicolas, E. RIFA, S. MONDY, M. Le Boulch, L. Cauquil, G. Hernandez Raquet, S. Terrat et A.-L. Abraham (2019). « Guide pratique à destination des biologistes, bioinformaticiens et statisticiens qui souhaitent s'initier aux analyses métabarcoding ». In : *Cahiers des Techniques de l'INRA* 2019.97, p. 1-23.

Federhen, S. (13 août 2003). *The Taxonomy Project*. National Center for Biotechnology Information (US).

Federhen, S. (1er jan. 2012). « The NCBI Taxonomy database ». In : *Nucleic Acids Research* 40 (D1), p. D136-D143. ISSN : 0305-1048. DOI : 10.1093/nar/gkr1178.

Federhen, S., K. Clark, T. Barrett, H. Parkinson, J. Ostell, Y. Kodama, J. Mashima, Y. Nakamura, G. Cochrane et I. Karsch-Mizrachi (nov. 2014). « Toward richer metadata for microbial sequences : replacing strain-level NCBI taxonomy taxids with BioProject, BioSample and Assembly records ». In : *Standards in Genomic Sciences* 9.3, p. 1275-1277. ISSN : 1944-3277. DOI : 10.4056/sigs.4851102.

Fierer, N., M. A. Bradford et R. B. Jackson (2007). « Toward an Ecological Classification of Soil Bacteria ». In : *Ecology* 88.6, p. 1354-1364. ISSN : 1939-9170. DOI : 10.1890/05-1839.

Fierer, N., J. Schimel et P. Holden (1er jan. 2003). « Influence of Drying–Rewetting Frequency on Soil Bacterial Community Structure ». In : *Microbial Ecology* 45.1, p. 63-71. issn : 1432-184X. doi : 10.1007/s00248-002-1007-2.

Figueras, M. J., R. Beaz-Hidalgo, M. J. Hossain et M. R. Liles (4 déc. 2014). « Taxonomic affiliation of new genomes should be verified using average nucleotide identity and multilocus phylogenetic analysis ». In : *Genome Announcements* 2.6. issn : 2169-8287. doi : 10.1128/genomeA.00927-14.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick et E. Al (28 juil. 1995). « Whole-genome random sequencing and assembly of Haemophilus influenzae Rd ». In : *Science* 269.5223, p. 496-512. issn : 0036-8075, 1095-9203. doi : 10.1126/science.7542800.

Fox, G. E., J. D. Wisotzkey et P. Jurtshuk (jan. 1992). « How close is close : 16S rRNA sequence identity may not be sufficient to guarantee species identity ». In : *International Journal of Systematic Bacteriology* 42.1, p. 166-170. issn : 0020-7713. doi : 10.1099/00207713-42-1-166.

Franzosa, E. A., L. J. McIver, G. Rahnavard, L. R. Thompson, M. Schirmer, G. Weingart, K. S. Lipson, R. Knight, J. G. Caporaso, N. Segata et C. Huttenhower (nov. 2018). « Species-level functional profiling of metagenomes and metatranscriptomes ». In : *Nature Methods* 15.11, p. 962. issn : 1548-7105. doi : 10.1038/s41592-018-0176-y.

Fraser, C., W. P. Hanage et B. G. Spratt (26 jan. 2007). « Recombination and the nature of bacterial speciation ». In : *Science (New York, N.Y.)* 315.5811, p. 476-480. issn : 0036-8075. doi : 10.1126/science.1127573.

Fraser, C. M., J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, J. L. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, J. M. Merrick, J.-F. Tomb, B. A. Dougherty, K. F. Bott, P.-C. Hu, T. S. Lucier, S. N. Peterson, H. O. Smith, C. A. Hutchison et J. C. Venter (20 oct. 1995). « The Minimal Gene Complement of Mycoplasma genitalium ». In : *Science* 270.5235, p. 397-404. issn : 0036-8075, 1095-9203. doi : 10.1126/science.270.5235.397.

Fuhrman, J. A. (13 mai 2009). « Microbial community structure and its functional implications ». In : *Nature* 459, p. 193-199. issn : 1476-4687. doi : 10.1038/nature08058.

Fuller, C. W., L. R. Middendorf, S. A. Benner, G. M. Church, T. Harris, X. Huang, S. B. Jovanovich, J. R. Nelson, J. A. Schloss, D. C. Schwartz et D. V. Vezenov (nov. 2009). « The challenges of sequencing by synthesis ». In : *Nature Biotechnology* 27.11, p. 1013-1023. issn : 1546-1696. doi : 10.1038/nbt.1585.

Garrity, G., éd. (2005). *Bergey's Manual of Systematic Bacteriology : Volume 2 : The Proteobacteria*. 2e éd. Bergey's Manual of Systematic Bacteriology. Springer US. isbn : 978-0-387-95040-2.

Garrity, G. M., J. A. Bell et T. G. Lilburn (2004). « Taxonomic outline of the prokaryotes. Bergey's manual of systematic bacteriology ». In : *Springer, New York, Berlin, Heidelberg*.

Gevers, D., F. M. Cohan, J. G. Lawrence, B. G. Spratt, T. Coenye, E. J. Feil, E. Stackebrandt, Y. V. d. Peer, P. Vandamme, F. L. Thompson et J. Swings (sept. 2005). « Re-evaluating prokaryotic species ». In : *Nature Reviews Microbiology* 3.9, p. 733. issn : 1740-1534. doi : 10.1038/nrmicro1236.

Gevers, D., P. Dawyndt, P. Vandamme, A. Willems, M. Vancanneyt, J. Swings et P. De Vos (29 nov. 2006). « Stepping stones towards a new prokaryotic taxonomy ». In : *Philosophical Transactions of the Royal Society B : Biological Sciences* 361.1475, p. 1911-1916. issn : 0962-8436. doi : 10.1098/rstb.2006.1915.

Gillott, C. (1995). « Taxonomy and Systematics ». In : *Entomology*. Dordrecht : Springer Netherlands, p. 91-112. isbn : 978-0-306-44967-3. doi : 10.1007/978-94-017-4380-8_4.

Glaeser, S. P. et P. Kämpfer (1er juin 2015). « Multilocus sequence analysis (MLSA) in prokaryotic taxonomy ». In : *Systematic and Applied Microbiology*. Taxonomy in the age of genomics 38.4, p. 237-245. issn : 0723-2020. doi : 10.1016/j.syapm.2015.03.007.

Gogarten, J. P. et J. P. Townsend (sept. 2005). « Horizontal gene transfer, genome innovation and evolution ». In : *Nature Reviews Microbiology* 3.9, p. 679. issn : 1740-1534. doi : 10.1038/nrmicro1204.

Goris, J., K. T. Konstantinidis, J. A. Klappenbach, T. Coenye, P. Vandamme et J. M. Tiedje (2007). « DNA–DNA hybridization values and their relationship to whole-genome sequence similarities ». In : *In-*

ternational Journal of Systematic and Evolutionary Microbiology 57.1, p. 81-91. DOI : 10.1099/ijs.0.64483-0.

GRIMONT, P. A. (avr. 1988). « Use of DNA reassociation in bacterial classification ». In : *Canadian Journal of Microbiology* 34.4, p. 541-546. ISSN : 0008-4166.

HALL, I. C. et E. O'TOOLE (1er fév. 1935). « Intestinal flora in newborn infants with a description of a new pathogenic anaerobe ». In : *American Journal of Diseases of Children* 49.2, p. 390-402. ISSN : 0096-8994. DOI : 10.1001/archpedi.1935.01970020105010.

HANNIFFY, S. B., C. PELÁEZ, M. A. MARTÍNEZ-BARTOLOMÉ, T. REQUENA et M. C. MARTÍNEZ-CUESTA (15 nov. 2009). « Key enzymes involved in methionine catabolism by cheese lactic acid bacteria ». In : *International Journal of Food Microbiology* 135.3, p. 223-230. ISSN : 0168-1605. DOI : 10.1016/j.ijfoodmicro.2009.08.009.

HASLAM, E. (1er jan. 1986). « Secondary metabolism – fact and fiction ». In : *Natural Product Reports* 3.0, p. 217-249. ISSN : 1460-4752. DOI : 10.1039/NP9860300217.

HELLER, S., A. MCNAUGHT, S. STEIN, D. TCHEKHOVSKOI et I. PLETNEV (24 jan. 2013). « InChI - the worldwide chemical structure identifier standard ». In : *Journal of Cheminformatics* 5.1, p. 7. ISSN : 1758-2946. DOI : 10.1186/1758-2946-5-7.

HENRIKSEN, S. D. (1er jan. 1978). « Chapter I Serotyping of Bacteria ». In : *Methods in Microbiology*. Sous la dir. de T. BERGAN et J. R. NORRIS. T. 12. Academic Press, p. 1-13. DOI : 10.1016/S0580-9517(08)70355-6.

HINCHLIFF, C. E., S. A. SMITH, J. F. ALLMAN, J. G. BURLEIGH, R. CHAUDHARY, L. M. COGHILL, K. A. CRANDALL, J. DENG, B. T. DREW, R. GAZIS, K. GUDE, D. S. HIBBETT, L. A. KATZ, H. D. LAUGHINGHOUSE, E. J. MCTAVISH, P. E. MIDFORD, C. L. OWEN, R. H. REE, J. A. REES, D. E. SOLTIS, T. WILLIAMS et K. A. CRANSTON (13 oct. 2015). « Synthesis of phylogeny and taxonomy into a comprehensive tree of life ». In : *Proceedings of the National Academy of Sciences* 112.41, p. 12764-12769. ISSN : 0027-8424, 1091-6490. DOI : 10.1073/pnas.1423041112.

HODGSON, D. A. (2000). « Primary metabolism and its control in streptomycetes : a most unusual group of bacteria ». In : *Advances in Microbial Physiology* 42, p. 47-238. ISSN : 0065-2911.

Holdeman, L. V., E. P. Cato et W. E. C. Moore (1977). « Anaerobic bacteriology manual ». In : *Anaerobe Laboratory, Virginia Polytechnic Institute and State University, Blacksburg*.

Huerta-Cepas, J., K. Forslund, L. P. Coelho, D. Szklarczyk, L. J. Jensen, C. von Mering et P. Bork (1er août 2017). « Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper ». In : *Molecular Biology and Evolution* 34.8, p. 2115-2122. issn : 0737-4038. doi : 10.1093/molbev/msx148.

Jain, C., S. Koren, A. Dilthey, A. M. Phillippy et S. Aluru (1er sept. 2018a). « A fast adaptive algorithm for computing whole-genome homology maps ». In : *Bioinformatics* 34.17, p. i748-i756. issn : 1367-4803. doi : 10.1093/bioinformatics/bty597.

Jain, C., L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis et S. Aluru (30 nov. 2018b). « High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries ». In : *Nature Communications* 9.1, p. 5114. issn : 2041-1723. doi : 10.1038/s41467-018-07641-9.

Jerphagnon, L. (1973). *Dictionnaire des Grandes Philosophies. Sous la Direction de Lucien Jerphagnon*. Privat.

Jeske, L., S. Placzek, I. Schomburg, A. Chang et D. Schomburg (8 jan. 2019). « BRENDA in 2019 : a European ELIXIR core data resource ». In : *Nucleic Acids Research* 47 (D1), p. D542-D549. issn : 0305-1048. doi : 10.1093/nar/gky1048.

Johnson, C. H., J. Ivanisevic, H. P. Benton et G. Siuzdak (6 jan. 2015). « Bioinformatics : The Next Frontier of Metabolomics ». In : *Analytical Chemistry* 87.1, p. 147-156. issn : 0003-2700. doi : 10.1021/ac5040693.

Jovel, J., J. Patterson, W. Wang, N. Hotte, S. O'Keefe, T. Mitchel, T. Perry, D. Kao, A. L. Mason, K. L. Madsen et G. K.-S. Wong (2016). « Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics ». In : *Frontiers in Microbiology* 7. issn : 1664-302X. doi : 10.3389/fmicb.2016.00459.

Kämpfer, P. et S. P. Glaeser (2012). « Prokaryotic taxonomy in the sequencing era – the polyphasic approach revisited ». In : *Environmental Microbiology* 14.2, p. 291-317. issn : 1462-2920. doi : 10.1111/j.1462-2920.2011.02615.x.

Kanehisa, M., M. Furumichi, M. Tanabe, Y. Sato et K. Morishima (4 jan. 2017). « KEGG : new perspectives on genomes, pathways, diseases

and drugs ». In : *Nucleic Acids Research* 45 (D1), p. D353-D361. ISSN : 0305-1048. DOI : 10.1093/nar/gkw1092.

KANEHISA, M. et S. GOTO (1er jan. 2000). « KEGG : Kyoto Encyclopedia of Genes and Genomes ». In : *Nucleic Acids Research* 28.1, p. 27-30. ISSN : 0305-1048.

KANEHISA, M., Y. SATO, M. FURUMICHI, K. MORISHIMA et M. TANABE (8 jan. 2019). « New approach for understanding genome variations in KEGG ». In : *Nucleic Acids Research* 47 (Database issue), p. D590-D595. ISSN : 0305-1048. DOI : 10.1093/nar/gky962.

KARP, P. D., R. BILLINGTON, R. CASPI, C. A. FULCHER, M. LATENDRESSE, A. KOTHARI, I. M. KESELER, M. KRUMMENACKER, P. E. MIDFORD, Q. ONG, W. K. ONG, S. M. PALEY et P. SUBHRAVETI (2017). « The BioCyc collection of microbial genomes and metabolic pathways ». In : *Briefings in Bioinformatics*. DOI : 10.1093/bib/bbx085.

KARP, P. D., N. IVANOVA, M. KRUMMENACKER, N. KYRPIDES, M. LATENDRESSE, P. MIDFORD, W. K. ONG, S. PALEY et R. SESHADRI (22 fév. 2019). « A Comparison of Microbial Genome Web Portals ». In : *Frontiers in Microbiology* 10. ISSN : 1664-302X. DOI : 10.3389/fmicb.2019.00208.

KARP, P. D., M. LATENDRESSE et R. CASPI (23 déc. 2011). « The Pathway Tools Pathway Prediction Algorithm ». In : *Standards in Genomic Sciences* 5.3, p. 424-429. ISSN : 1944-3277. DOI : 10.4056/sigs.1794338.

KARP, P. D., M. LATENDRESSE, S. M. PALEY, M. KRUMMENACKER, Q. D. ONG, R. BILLINGTON, A. KOTHARI, D. WEAVER, T. LEE, P. SUBHRAVETI, A. SPAULDING, C. FULCHER, I. M. KESELER et R. CASPI (sept. 2016). « Pathway Tools version 19.0 update : software for pathway/genome informatics and systems biology ». In : *Briefings in Bioinformatics* 17.5, p. 877-890. ISSN : 1477-4054. DOI : 10.1093/bib/bbv079.

KARP, P. D., M. LATENDRESSE, S. M. PALEY, M. K. Q. ONG, R. BILLINGTON, A. KOTHARI, D. WEAVER, T. LEE, P. SUBHRAVETI, A. SPAULDING, C. FULCHER, I. M. KESELER et R. CASPI (14 oct. 2015). « Pathway Tools version 19.0 : Integrated Software for Pathway/Genome Informatics and Systems Biology ». In : *arXiv :1510.03964 [q-bio]*. arXiv : 1510.03964.

KARP, P. D., M. RILEY, M. SAIER, I. T. PAULSEN, S. M. PALEY et A. PELLEGRINI-TOOLE (1er jan. 2000). « The EcoCyc and MetaCyc databases ». In : *Nucleic Acids Research* 28.1, p. 56-59. ISSN : 0305-1048.

KARP, P. D., D. WEAVER, S. PALEY, C. FULCHER, A. KUBO, A. KOTHARI, M. KRUMMENACKER, P. SUBHRAVETI, D. WEERASINGHE, S. GAMA-

Castro, A. M. Huerta, L. Muñiz-Rascado, C. Bonavides-Martinez, V. Weiss, M. Peralta-Gil, A. Santos-Zavaleta, I. Schröder, A. Mackie, R. Gunsalus, J. Collado-Vides, I. M. Keseler et I. Paulsen (21 mar. 2014). « The EcoCyc Database ». In : *EcoSal Plus* 6.1. issn : 2324-6200. doi : 10.1128/ecosalplus.ESP-0009-2013.

Kent, W. J. (1er avr. 2002). « BLAT—The BLAST-Like Alignment Tool ». In : *Genome Research* 12.4, p. 656-664. issn : 1088-9051, 1549-5469. doi : 10.1101/gr.229202.

Kersters, K., B. Pot, D. Dewettinck, U. Torck, M. Vancanneyt, L. Vauterin et P. Vandamme (1994). « Identification and Typing of Bacteria by Protein Electrophoresis ». In : *Bacterial Diversity and Systematics*. Sous la dir. de F. G. Priest, A. Ramos-Cormenzana et B. J. Tindall. Federation of European Microbiological Societies Symposium Series. Boston, MA : Springer US, p. 51-66. isbn : 978-1-4615-1869-3. doi : 10.1007/978-1-4615-1869-3_3.

Keseler, I. M., A. Mackie, M. Peralta-Gil, A. Santos-Zavaleta, S. Gama-Castro, C. Bonavides-Martínez, C. Fulcher, A. M. Huerta, A. Kothari, M. Krummenacker, M. Latendresse, L. Muñiz-Rascado, Q. Ong, S. Paley, I. Schröder, A. G. Shearer, P. Subhraveti, M. Travers, D. Weerasinghe, V. Weiss, J. Collado-Vides, R. P. Gunsalus, I. Paulsen et P. D. Karp (1er jan. 2013). « EcoCyc : fusing model organism databases with systems biology ». In : *Nucleic Acids Research* 41 (D1), p. D605-D612. issn : 0305-1048, 1362-4962. doi : 10.1093/nar/gks1027.

Kielak, A. M., C. C. Barreto, G. A. Kowalchuk, V. Veen, J. A et E. E. Kuramae (2016). « The Ecology of Acidobacteria : Moving beyond Genes and Genomes ». In : *Frontiers in Microbiology* 7. issn : 1664-302X. doi : 10.3389/fmicb.2016.00744.

Kitahara, K. et K. Miyazaki (1er jan. 2013). « Revisiting bacterial phylogeny ». In : *Mobile Genetic Elements* 3.1. issn : 2159-2543. doi : 10.4161/mge.24210.

Klingenberg, H., K. P. Asshauer, T. Lingner et P. Meinicke (15 avr. 2013). « Protein signature-based estimation of metagenomic abundances including all domains of life and viruses ». In : *Bioinformatics* 29.8, p. 973-980. issn : 1367-4803. doi : 10.1093/bioinformatics/btt077.

Ko, C. Y., J. L. Johnson, L. B. Barnett, H. M. McNair et J. R. Vercellotti (15 mai 1977). « A sensitive estimation of the percentage of guanine plus cytosine in deoxyribonucleic acid by high performance

liquid chromatography ». In : *Analytical Biochemistry* 80.1, p. 183-192. ISSN : 0003-2697. DOI : 10.1016/0003-2697(77)90638-8.

KONSTANTINIDIS, K. T. et J. M. TIEDJE (15 fév. 2005a). « Genomic insights that advance the species definition for prokaryotes ». In : *Proceedings of the National Academy of Sciences of the United States of America* 102.7, p. 2567-2572. ISSN : 0027-8424. DOI : 10.1073/pnas.0409727102.

KONSTANTINIDIS, K. T. et J. M. TIEDJE (sept. 2005b). « Towards a Genome-Based Taxonomy for Prokaryotes ». In : *Journal of Bacteriology* 187.18, p. 6258-6264. ISSN : 0021-9193. DOI : 10.1128/JB.187.18.6258-6264.2005.

KOONIN, E. V. et M. Y. GALPERIN (2003). « Evolution of Central Metabolic Pathways : The Playground of Non-orthologous Gene Displacement ». In : *Sequence — Evolution — Function*. Boston, MA : Springer US, p. 295-355. ISBN : 978-1-4419-5321-6. DOI : 10.1007/978-1-4757-3783-7_8.

KRIEG, N. R., W. LUDWIG, W. WHITMAN, B. P. HEDLUND, B. J. PASTER, J. T. STALEY, N. WARD, D. BROWN et A. PARTE, éd. (2010). *Bergey's Manual of Systematic Bacteriology : Volume 4 : The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes*. 2e éd. Bergey's Manual of Systematic Bacteriology. New York : Springer-Verlag. ISBN : 978-0-387-95042-6.

KRUMMENACKER, M., S. PALEY, L. MUELLER, T. YAN et P. D. KARP (15 août 2005). « Querying and computing with BioCyc databases ». In : *Bioinformatics* 21.16, p. 3454-3455. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/bti546.

KURTZ, S., A. PHILLIPPY, A. L. DELCHER, M. SMOOT, M. SHUMWAY, C. ANTONESCU et S. L. SALZBERG (30 jan. 2004). « Versatile and open software for comparing large genomes ». In : *Genome Biology* 5.2, R12. ISSN : 1474-760X. DOI : 10.1186/gb-2004-5-2-r12.

LA REAU, A. J., J. P. MEIER-KOLTHOFF et G. SUEN (2016). « Sequence-based analysis of the genus Ruminococcus resolves its phylogeny and reveals strong host association ». In : *Microbial Genomics* 2.12. DOI : 10.1099/mgen.0.000099.

LAND, M., L. HAUSER, S.-R. JUN, I. NOOKAEW, M. R. LEUZE, T.-H. AHN, T. KARPINETS, O. LUND, G. KORA, T. WASSENAAR, S. POUDEL et D. W. USSERY (2015). « Insights from 20 years of bacterial genome sequencing ». In : *Functional & Integrative Genomics* 15.2, p. 141-161. ISSN : 1438-793X. DOI : 10.1007/s10142-015-0433-4.

Langille, M. G. I., J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. Vega Thurber, R. Knight, R. G. Beiko et C. Huttenhower (sept. 2013). « Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences ». In : *Nature Biotechnology* 31.9, p. 814-821. issn : 1087-0156. doi : 10.1038/nbt.2676.

Langmead, B. et S. L. Salzberg (avr. 2012). « Fast gapped-read alignment with Bowtie 2 ». In : *Nature Methods* 9.4, p. 357-359. issn : 1548-7105. doi : 10.1038/nmeth.1923.

Lapage, S. P., P. H. A. Sneath, E. F. Lessel, V. B. D. Skerman, H. P. R. Seeliger et W. A. Clark, éd. (1992). *International Code of Nomenclature of Bacteria : Bacteriological Code, 1990 Revision*. Washington (DC) : ASM Press. isbn : 978-1-55581-039-9.

Larocque, M., T. Chénard et R. Najmanovich (15 oct. 2014). « A curated C. difficile strain 630 metabolic network : prediction of essential targets and inhibitors ». In : *BMC Systems Biology* 8.1, p. 117. issn : 1752-0509. doi : 10.1186/s12918-014-0117-z.

Lavergne, C. (11 déc. 2014). « Rôle (structure et fonction) des communautés procaryotes (bactéries et archées) dans le cycle de l'azote d'une vasière littorale du Pertuis Charentais : influence des facteurs biotiques et abiotiques par une approche multi-échelle ». thesis. La Rochelle.

Lawrence, J. G. (1er juin 2002). « Gene Transfer in Bacteria : Speciation without Species ? » In : *Theoretical Population Biology* 61.4, p. 449-460. issn : 0040-5809. doi : 10.1006/tpbi.2002.1587.

Lawson, P. A., D. M. Citron, K. L. Tyrrell et S. M. Finegold (1er août 2016). « Reclassification of Clostridium difficile as Clostridioides difficile (Hall and O'Toole 1935) Prévot 1938 ». In : *Anaerobe* 40, p. 95-99. issn : 1075-9964. doi : 10.1016/j.anaerobe.2016.06.008.

Lawson, P. A. et S. M. Finegold (2015). « Reclassification of Ruminococcus obeum as Blautia obeum comb. nov. » In : *International Journal of Systematic and Evolutionary Microbiology* 65.3, p. 789-793. doi : 10.1099/ijs.0.000015.

Le Boulch, M., S. Combes et G. Pascal (16 mai 2017). « Functional inference of complex bacterial communities from marker genes derived from high-throughput sequencing ». Présentation orale. Présentation orale. NEM 2017. Saint Pée-sur-Nivelle.

Le Boulch, M., S. Combes et G. Pascal (16 mai 2018). « Difficultés de l'inférence fonctionnelle à partir du 16S et présentation de MACADAM ». Présentation orale. Présentation orale. NEM 2018. Narbonne.

Le Boulch, M., P. Dehais, S. Combes et G. Pascal (7 avr. 2018a). « MACADAM a user-friendly MetAboliC pAthway DAtabase for complex Microbial community function analysis ». Présentation orale. Présentation orale. JOBIM 2018. Marseille.

Le Boulch, M., P. Dehais, S. Combes et G. Pascal (11 sept. 2018b). « MACADAM : a user-friendly MetAboliC pAthway DAtabase for complex Microbial community function analysis ». Poster. Poster. ECCB 2018. Athènes.

Le Boulch, M., P. Déhais, S. Combes et G. Pascal (2018c). *MACADAMExplore*. URL : http://macadam.toulouse.inra.fr/.

Le Boulch, M., P. Déhais, S. Combes et G. Pascal (1er jan. 2019). « The MACADAM database : a MetAboliC pAthways DAtabase for Microbial taxonomic groups for mining potential metabolic capacities of archaeal and bacterial taxonomic groups ». In : *Database* 2019. DOI : 10.1093/database/baz049.

Lee, I., Y. Ouk Kim, S.-C. Park et J. Chun (2016). « OrthoANI : An improved algorithm and software for calculating average nucleotide identity ». In : *International Journal of Systematic and Evolutionary Microbiology* 66.2, p. 1100-1103. DOI : 10.1099/ijsem.0.000760.

Leeuwenhoek, A. V. (25 mar. 1677). « Observations, communicated to the publisher by Mr. Antony van Leewenhoeck, in a dutch letter of the 9th Octob. 1676. here English'd : concerning little animals by him observed in rain-well-sea- and snow water ; as also in water wherein pepper had lain infused ». In : *Philosophical Transactions of the Royal Society of London* 12.133, p. 821-831. DOI : 10.1098/rstl.1677.0003.

Leigh, J. (25 nov. 2011). *Principles of Chemical Nomenclature*. ISBN : 978-1-84973-007-5.

Leinonen, R., H. Sugawara et M. Shumway (jan. 2011). « The Sequence Read Archive ». In : *Nucleic Acids Research* 39 (Database issue), p. D19-D21. ISSN : 0305-1048. DOI : 10.1093/nar/gkq1019.

Lengauer, T. et C. Hartmann (1er jan. 2007). « 3.15 - Bioinformatics ». In : *Comprehensive Medicinal Chemistry II*. Sous la dir. de J. B. Taylor et D. J. Triggle. Oxford : Elsevier, p. 315-347. ISBN : 978-0-08-045044-5. DOI : 10.1016/B0-08-045044-X/00088-2.

Ley, J. D., H. Cattoir et A. Reynaerts (1970). « The Quantitative Measurement of DNA Hybridization from Renaturation Rates ». In : *European Journal of Biochemistry* 12.1, p. 133-142. ISSN : 1432-1033. DOI : 10.1111/j.1432-1033.1970.tb00830.x.

Li, D., R. Luo, C.-M. Liu, C.-M. Leung, H.-F. Ting, K. Sadakane, H. Yamashita et T.-W. Lam (1er juin 2016). « MEGAHIT v1.0 : A fast and scalable metagenome assembler driven by advanced methodologies and community practices ». In : *Methods*. Pan-omics analysis of biological data 102, p. 3-11. ISSN : 1046-2023. DOI : 10.1016/j.ymeth.2016.02.020.

Linné, C. v. et L. Salvius (1758). *Systema naturae per regna tria naturae :secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis.* Holmiae : Impensis Direct. Laurentii Salvii, 881 p.

Liu, C., S. M. Finegold, Y. Song et P. A. Lawson (2008). « Reclassification of Clostridium coccoides, Ruminococcus hansenii, Ruminococcus hydrogenotrophicus, Ruminococcus luti, Ruminococcus productus and Ruminococcus schinkii as Blautia coccoides gen. nov., comb. nov., Blautia hansenii comb. nov., Blautia hydrogenotrophica comb. nov., Blautia luti comb. nov., Blautia producta comb. nov., Blautia schinkii comb. nov. and description of Blautia wexlerae sp. nov., isolated from human faeces ». In : *International Journal of Systematic and Evolutionary Microbiology* 58.8, p. 1896-1902. DOI : 10.1099/ijs.0.65208-0.

Liu, K., C. R. Linder et T. Warnow (21 nov. 2011). « RAxML and FastTree : Comparing Two Methods for Large-Scale Maximum Likelihood Phylogeny Estimation ». In : *PLOS ONE* 6.11, e27731. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0027731.

Louca, S. et M. Doebeli (15 mar. 2018). « Efficient comparative phylogenetics on large trees ». In : *Bioinformatics* 34.6, p. 1053-1055. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btx701.

Louca, S., L. W. Parfrey et M. Doebeli (16 sept. 2016). « Decoupling function and taxonomy in the global ocean microbiome ». In : *Science* 353.6305, p. 1272-1277. ISSN : 0036-8075, 1095-9203. DOI : 10.1126/science.aaf4507.

Ludwig, W., O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Förster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. König, T. Liss, R. Lüssmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig,

M. Lenke, T. Ludwig, A. Bode et K.-H. Schleifer (2004). « ARB : a software environment for sequence data ». In : *Nucleic Acids Research* 32.4, p. 1363-1371. issn : 0305-1048. doi : 10.1093/nar/gkh293.

Mahé, F., T. Rognes, C. Quince, C. d. Vargas et M. Dunthorn (10 déc. 2015). « Swarm v2 : highly-scalable and high-resolution amplicon clustering ». In : *PeerJ* 3, e1420. issn : 2167-8359. doi : 10.7717/peerj.1420.

Mahé, F., T. Rognes, C. Quince, C. de Vargas et M. Dunthorn (25 sept. 2014). « Swarm : robust and fast clustering method for amplicon-based studies ». In : *PeerJ* 2. issn : 2167-8359. doi : 10.7717/peerj.593.

Maiden, M. C. J., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman et B. G. Spratt (17 mar. 1998). « Multilocus sequence typing : A portable approach to the identification of clones within populations of pathogenic microorganisms ». In : *Proceedings of the National Academy of Sciences* 95.6, p. 3140-3145. issn : 0027-8424, 1091-6490. doi : 10.1073/pnas.95.6.3140.

Maki, J. J. et T. Looft (nov. 2018). « Megasphaera stantonii sp. nov., a butyrate-producing bacterium isolated from the cecum of a healthy chicken ». In : *International Journal of Systematic and Evolutionary Microbiology* 68.11, p. 3409-3415. issn : 1466-5034. doi : 10.1099/ijsem.0.002991.

Marchandin, H., C. Teyssier, M. Siméon de Buochberg, H. Jean-Pierre, C. Carriere et E. Jumas-Bilak (2003). « Intra-chromosomal heterogeneity between the four 16S rRNA gene copies in the genus Veillonella : implications for phylogeny and taxonomy ». In : *Microbiology* 149.6, p. 1493-1501. doi : 10.1099/mic.0.26132-0.

Markley, J. L., R. Brüschweiler, A. S. Edison, H. R. Eghbalnia, R. Powers, D. Raftery et D. S. Wishart (1er fév. 2017). « The future of NMR-based metabolomics ». In : *Current Opinion in Biotechnology*. Analytical biotechnology 43, p. 34-40. issn : 0958-1669. doi : 10.1016/j.copbio.2016.08.001.

Markowitz, V. M., I.-M. A. Chen, K. Chu, E. Szeto, K. Palaniappan, Y. Grechkin, A. Ratner, B. Jacob, A. Pati, M. Huntemann, K. Liolios, I. Pagani, I. Anderson, K. Mavromatis, N. N. Ivanova et N. C. Kyrpides (jan. 2012). « IMG/M : the integrated metagenome data management and comparative analysis system ». In : *Nucleic Acids Research* 40 (Database issue), p. D123-D129. issn : 0305-1048. doi : 10.1093/nar/gkr975.

Marmur, J. et P. Doty (1er juil. 1962). « Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature ». In : *Journal of Molecular Biology* 5.1, p. 109-118. issn : 0022-2836. doi : 10.1016/S0022-2836(62)80066-7.

Mayr, E. (1942). *Systematics and the Origin of Species*. Columbia Univ. Press.

Mayr, E. (1er mar. 1966). « The Proper Spelling of Taxonomy ». In : *Systematic Biology* 15.1, p. 88-88. issn : 1063-5157. doi : 10.2307/sysbio/15.1.88a.

McCutcheon, J. P. et N. A. Moran (1er jan. 2010). « Functional Convergence in Reduced Genomes of Bacterial Symbionts Spanning 200 My of Evolution ». In : *Genome Biology and Evolution* 2, p. 708-718. doi : 10.1093/gbe/evq055.

McDonald, A. G., S. Boyce, G. P. Moss, H. B. F. Dixon et K. F. Tipton (27 juil. 2007). « ExplorEnz : a MySQL database of the IUBMB enzyme nomenclature ». In : *BMC biochemistry* 8, p. 14. issn : 1471-2091. doi : 10.1186/1471-2091-8-14.

McDonald, A. G., S. Boyce et K. F. Tipton (2015). « Enzyme Classification and Nomenclature ». In : *eLS*. American Cancer Society, p. 1-11. isbn : 978-0-470-01590-2. doi : 10.1002/9780470015902.a0000710.pub3.

McDonald, A. G., S. Boyce et K. F. Tipton (jan. 2009). « ExplorEnz : the primary source of the IUBMB enzyme list ». In : *Nucleic Acids Research* 37 (Database issue), p. D593-597. issn : 1362-4962. doi : 10.1093/nar/gkn582.

McDonald, D., M. N. Price, J. Goodrich, E. P. Nawrocki, T. Z. DeSantis, A. Probst, G. L. Andersen, R. Knight et P. Hugenholtz (mar. 2012). « An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea ». In : *The ISME Journal* 6.3, p. 610-618. issn : 1751-7362. doi : 10.1038/ismej.2011.139.

Medini, D., C. Donati, H. Tettelin, V. Masignani et R. Rappuoli (1er déc. 2005). « The microbial pan-genome ». In : *Current Opinion in Genetics & Development*. Genomes and evolution 15.6, p. 589-594. issn : 0959-437X. doi : 10.1016/j.gde.2005.09.006.

Meier-Kolthoff, J. P., H.-P. Klenk et M. Göker (2014). « Taxonomic use of DNA G+C content and DNA–DNA hybridization in the genomic

age ». In : *International Journal of Systematic and Evolutionary Microbiology* 64.2, p. 352-356. DOI : 10.1099/ijs.0.056994-0.

Meier, R., C. Ruttkies, H. Treutler et S. Neumann (10 nov. 2017). « Bioinformatics can boost metabolomics research ». In : *Journal of Biotechnology*. Bioinformatics Solutions for Big Data Analysis in Life Sciences presented by the German Network for Bioinformatics Infrastructure 261, p. 137-141. ISSN : 0168-1656. DOI : 10.1016/j.jbiotec.2017.05.018.

Meinicke, P. (1er mai 2015). « UProC : tools for ultra-fast protein domain classification ». In : *Bioinformatics* 31.9, p. 1382-1388. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btu843.

Mercier, C., F. Boyer, A. Bonin et E. Coissac (2013). « SUMATRA and SUMACLUST : fast and exact comparison and clustering of sequences ». In : *Programs and Abstracts of the SeqBio 2013 workshop. Abstract*. Citeseer, p. 27-29.

Mesbah, N. M., W. B. Whitman et M. Mesbah (1er jan. 2011). « 14 - Determination of the G+C Content of Prokaryotes ». In : *Methods in Microbiology*. Sous la dir. de F. Rainey et A. Oren. T. 38. Taxonomy of Prokaryotes. Academic Press, p. 299-324. DOI : 10.1016/B978-0-12-387730-7.00014-0.

Methé, B. A., K. E. Nelson, M. Pop, H. H. Creasy, M. G. Giglio, C. Huttenhower, D. Gevers, J. F. Petrosino, S. Abubucker et J. H. Badger (13 juin 2012). « A framework for human microbiome research ». In : *Nature* 486.7402, p. 215-221.

Mignard, S. et J. P. Flandrois (1er déc. 2006). « 16S rRNA sequencing in routine bacterial identification : A 30-month experiment ». In : *Journal of Microbiological Methods* 67.3, p. 574-581. ISSN : 0167-7012. DOI : 10.1016/j.mimet.2006.05.009.

Moore, E. R. B., S. A. Mihaylova, P. Vandamme, M. I. Krichevsky et L. Dijkshoorn (1er juil. 2010). « Microbial systematics and taxonomy : relevance for a microbial commons ». In : *Research in Microbiology*. Microbial research commons : From strain isolation to practical use 161.6, p. 430-438. ISSN : 0923-2508. DOI : 10.1016/j.resmic.2010.05.007.

Moreira, A. P. B., N. Pereira et F. L. Thompson (2011). « Usefulness of a real-time PCR platform for G+C content and DNA–DNA hybridization estimations in vibrios ». In : *International Journal of Systematic and Evolutionary Microbiology* 61.10, p. 2379-2383. DOI : 10.1099/ijs.0.023606-0.

Morgat, A., E. Coissac, E. Coudert, K. B. Axelsen, G. Keller, A. Bairoch, A. Bridge, L. Bougueleret, I. Xenarios et A. Viari (1er jan. 2012). « UniPathway : a resource for the exploration and annotation of metabolic pathways ». In : *Nucleic Acids Research* 40 (D1), p. D761-D769. issn : 0305-1048. doi : 10.1093/nar/gkr1023.

Morrison, D. J. et T. Preston (3 mai 2016). « Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism ». In : *Gut Microbes* 7.3, p. 189-200. issn : 1949-0976. doi : 10.1080/19490976.2015.1134082.

Müller, S., B. Appel, D. Balke, R. Hieronymus et C. Nübel (27 juin 2016). « Thirty-five years of research into ribozymes and nucleic acid catalysis : where do we stand today ? » In : *F1000Research* 5. issn : 2046-1402. doi : 10.12688/f1000research.8601.1.

Munoz, R., P. Yarza, W. Ludwig, J. Euzéby, R. Amann, K.-H. Schleifer, F. Oliver Glöckner et R. Rosselló-Móra (1er mai 2011). « Release LTPs104 of the All-Species Living Tree ». In : *Systematic and Applied Microbiology* 34.3, p. 169-170. issn : 0723-2020. doi : 10.1016/j.syapm.2011.03.001.

Murray, R. G. et K. H. Schleifer (jan. 1994). « Taxonomic notes : a proposal for recording the properties of putative taxa of procaryotes ». In : *International Journal of Systematic Bacteriology* 44.1, p. 174-176. issn : 0020-7713. doi : 10.1099/00207713-44-1-174.

Muto, A. et S. Osawa (jan. 1987). « The guanine and cytosine content of genomic DNA and bacterial evolution. » In : *Proceedings of the National Academy of Sciences of the United States of America* 84.1, p. 166-169. issn : 0027-8424.

Mysara, M., P. Vandamme, R. Props, F.-M. Kerckhof, N. Leys, N. Boon, J. Raes et P. Monsieurs (1er avr. 2017). « Reconciliation between operational taxonomic units and species boundaries ». In : *FEMS Microbiology Ecology* 93.4. issn : 0168-6496. doi : 10.1093/femsec/fix029.

Neet, K. E. (2 oct. 1998). « Enzyme Catalytic Power Minireview Series ». In : *Journal of Biological Chemistry* 273.40, p. 25527-25528. issn : 0021-9258, 1083-351X. doi : 10.1074/jbc.273.40.25527.

Nguyen, N.-P., T. Warnow, M. Pop et B. White (20 avr. 2016). « A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity ». In : *npj Biofilms and Microbiomes* 2, p. 16004. issn : 2055-5008. doi : 10.1038/npjbiofilms.2016.4.

Nielsen, J. (20 juin 2017). « Systems Biology of Metabolism ». In : *Annual Review of Biochemistry* 86.1, p. 245-275. issn : 0066-4154. doi : 10.1146/annurev-biochem-061516-044757.

Noor, E., E. Eden, R. Milo et U. Alon (10 sept. 2010). « Central Carbon Metabolism as a Minimal Biochemical Walk between Precursors for Biomass and Energy ». In : *Molecular Cell* 39.5, p. 809-820. issn : 1097-2765. doi : 10.1016/j.molcel.2010.08.031.

Nurk, S., D. Meleshko, A. Korobeynikov et P. A. Pevzner (mai 2017). « metaSPAdes : a new versatile metagenomic assembler ». In : *Genome Research* 27.5, p. 824-834. issn : 1088-9051. doi : 10.1101/gr.213959.116.

O'Malley, M. A., A. G. B. Simpson et A. J. Roger (1er mar. 2013). « The other eukaryotes in light of evolutionary protistology ». In : *Biology & Philosophy* 28.2, p. 299-330. issn : 1572-8404. doi : 10.1007/s10539-012-9354-y.

Olsen, G. J., D. J. Lane, S. J. Giovannoni, N. R. Pace et D. A. Stahl (1er oct. 1986). « Microbial Ecology and Evolution : A Ribosomal RNA Approach ». In : *Annual Review of Microbiology* 40.1, p. 337-365. issn : 0066-4227. doi : 10.1146/annurev.mi.40.100186.002005.

Olsen, G. J., R. Overbeek, N. Larsen, T. L. Marsh, M. J. McCaughey, M. A. Maciukenas, W.-M. Kuan, T. J. Macke, Y. Xing et C. R. Woese (11 mai 1992). « The Ribosomal Database Project ». In : *Nucleic Acids Research* 20 (suppl), p. 2199-2200. issn : 0305-1048. doi : 10.1093/nar/20.suppl.2199.

Oren, A. (2009). « Metabolic Diversity in Prokaryotes and Eukaryotes ». In : *Biological Science Fundamentals and Systematics- Volume II*. Sous la dir. de G. Contrafatto et A. Minelli. EOLSS Publications. isbn : 978-1-84826-305-5.

Oren, A. (2011). « Cyanobacterial systematics and nomenclature as featured in the International Bulletin of Bacteriological Nomenclature and Taxonomy / International Journal of Systematic Bacteriology / International Journal of Systematic and Evolutionary Microbiology ». In : *International Journal of Systematic and Evolutionary Microbiology* 61.1, p. 10-15. doi : 10.1099/ijs.0.018838-0.

Oren, A., M. S. da Costa, G. M. Garrity, F. A. Rainey, R. Rosselló-Móra, B. Schink, I. Sutcliffe, M. E. Trujillo et W. B. Whitman (2015). « Proposal to include the rank of phylum in the International Code of Nomenclature of Prokaryotes ». In : *International Journal of*

Systematic and Evolutionary Microbiology, 65.11, p. 4284-4287. ISSN : 1466-5026, DOI : 10.1099/ijsem.0.000664.

OREN, A. et G. M. GARRITY (2016). « List of new names and new combinations previously effectively, but not validly, published ». In : *International Journal of Systematic and Evolutionary Microbiology* 66.9, p. 3761-3764. DOI : 10.1099/ijsem.0.001321.

OREN, A. et G. M. GARRITY (14 nov. 2018a). « Notification that new names of prokaryotes, new combinations, and new taxonomic opinions have appeared in volume 68, part 10 of the IJSEM ». In : *International Journal of Systematic and Evolutionary Microbiology,* 69.1, p. 10-12. ISSN : 1466-5026, DOI : 10.1099/ijsem.0.003107.

OREN, A. et G. M. GARRITY (7 fév. 2018b). « Notification that new names of prokaryotes, new combinations, and new taxonomic opinions have appeared in volume 68, part 4, of the IJSEM ». In : *International Journal of Systematic and Evolutionary Microbiology* 68.7, p. 2134-2136. DOI : 10.1099/ijsem.0.002764.

OREN, A. et G. M. GARRITY (4 jan. 2019a). « List of new names and new combinations previously effectively, but not validly, published ». In : *International Journal of Systematic and Evolutionary Microbiology* 69.1, p. 5-9. DOI : 10.1099/ijsem.0.003174.

OREN, A. et G. M. GARRITY (24 jan. 2019b). « Notification that new names of prokaryotes, new combinations, and new taxonomic opinions have appeared in volume 69, part 1, of the IJSEM ». In : *International Journal of Systematic and Evolutionary Microbiology* 69.4, p. 875-876. DOI : 10.1099/ijsem.0.003251.

OREN, A. et S. VENTURA (1er oct. 2017). « The current status of cyanobacterial nomenclature under the "prokaryotic" and the "botanical" code ». In : *Antonie van Leeuwenhoek* 110.10, p. 1257-1269. ISSN : 1572-9699. DOI : 10.1007/s10482-017-0848-0.

OUZOUNIS, C. et N. KYRPIDES (1996). « The emergence of major cellular processes in evolution ». In : *FEBS Letters* 390.2, p. 119-123. ISSN : 1873-3468. DOI : 10.1016/0014-5793(96)00631-X.

OWEN, R. J., L. R. HILL et S. P. LAPAGE (1969). « Determination of DNA base compositions from melting profiles in dilute buffers ». In : *Biopolymers* 7.4, p. 503-516. ISSN : 1097-0282. DOI : 10.1002/bip.1969.360070408.

Pace, N. R., D. A. Stahl, D. J. Lane et G. J. Olsen (1986). « The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences ». In : *Advances in Microbial Ecology*. Sous la dir. de K. C. Marshall. Advances in Microbial Ecology. Boston, MA : Springer US, p. 1-55. isbn : 978-1-4757-0611-6. doi : 10.1007/978-1-4757-0611-6_1.

Paek, J., Y. Shin, J.-K. Kook et Y.-H. Chang (2019). « Blautia argi sp. nov., a new anaerobic bacterium isolated from dog faeces ». In : *International Journal of Systematic and Evolutionary Microbiology* 69.1, p. 33-38. doi : 10.1099/ijsem.0.002981.

Paley, S. M., M. Latendresse et P. D. Karp (24 sept. 2012). « Regulatory network operations in the Pathway Tools software ». In : *BMC Bioinformatics* 13.1, p. 243. issn : 1471-2105. doi : 10.1186/1471-2105-13-243.

Palinska, K. A. et W. Surosz (1er nov. 2014). « Taxonomy of cyanobacteria : a contribution to consensus approach ». In : *Hydrobiologia* 740.1, p. 1-11. issn : 1573-5117. doi : 10.1007/s10750-014-1971-9.

Palmer, M., S. N. Venter, M. P. A. Coetzee et E. T. Steenkamp (1er mar. 2019). « Prokaryotic species are sui generis evolutionary units ». In : *Systematic and Applied Microbiology* 42.2, p. 145-158. issn : 0723-2020. doi : 10.1016/j.syapm.2018.10.002.

Parker, C. T., B. J. Tindall et G. M. Garrity (1er jan. 2019). « International Code of Nomenclature of Prokaryotes : Prokaryotic Code (2008 Revision) ». In : *International Journal of Systematic and Evolutionary Microbiology* 69.1, S1-S111. issn : 1466-5026, 1466-5034. doi : 10.1099/ijsem.0.000778.

Parkhill, J., E. Birney et P. Kersey (2010). « Genomic information infrastructure after the deluge ». In : *Genome Biology* 11.7, p. 402. issn : 1465-6906. doi : 10.1186/gb-2010-11-7-402.

Parks, D. H., M. Chuvochina, P.-A. Chaumeil, C. Rinke, A. J. Mussig et P. Hugenholtz (18 sept. 2019). « Selection of representative genomes for 24,706 bacterial and archaeal species clusters provide a complete genome-based taxonomy ». In : *bioRxiv*, p. 771964. doi : 10.1101/771964.

Parks, D. H., M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil et P. Hugenholtz (oct. 2018). « A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life ». In : *Nature Biotechnology* 36.10, p. 996-1004. issn : 1546-1696. doi : 10.1038/nbt.4229.

Parte, A. C. (1er jan. 2014). « LPSN - list of prokaryotic names with standing in nomenclature ». In : *Nucleic Acids Research* 42 (Database issue), p. D613-D616. ISSN : 0305-1048. DOI : 10.1093/nar/gkt1111.

Parte, A. C. (2018). « LPSN - List of Prokaryotic names with Standing in Nomenclature (bacterio.net), 20 years on ». In : *International Journal of Systematic and Evolutionary Microbiology* 68.6, p. 1825-1829. DOI : 10.1099/ijsem.0.002786.

Pasolli, E., F. Asnicar, S. Manara, M. Zolfo, N. Karcher, F. Armanini, F. Beghini, P. Manghi, A. Tett, P. Ghensi, M. C. Collado, B. L. Rice, C. DuLong, X. C. Morgan, C. D. Golden, C. Quince, C. Huttenhower et N. Segata (24 jan. 2019). « Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle ». In : *Cell* 176.3, 649-662.e20. ISSN : 0092-8674, 1097-4172. DOI : 10.1016/j.cell.2019.01.001.

Paul, J. H. (août 1999). « Microbial gene transfer : an ecological perspective ». In : *Journal of Molecular Microbiology and Biotechnology* 1.1, p. 45-50. ISSN : 1464-1801.

Pham, T.-P.-T., F. Cadoret, M. Alou, S. Brah, B. Diallo, A. Diallo, C. Sokhna, J. Delerce, P.-E. Fournier, M. Million et D. Raoult (3 mar. 2017a). « 'Urmitella timonensis' gen. nov., sp. nov., 'Blautia marasmi' sp. nov., 'Lachnoclostridium pacaense' sp. nov., 'Bacillus marasmi' sp. nov. and 'Anaerotruncus rubiinfantis' sp. nov., isolated from stool samples of undernourished African children ». In : *New Microbes and New Infections* 17, p. 84-88. ISSN : 2052-2975. DOI : 10.1016/j.nmni.2017.02.004.

Pham, T.-P.-T., F. Cadoret, M. Tidjani Alou, S. Brah, B. Ali Diallo, A. Diallo, C. Sokhna, J. Delerce, P.-E. Fournier, M. Million et D. Raoult (12 mai 2017b). « 'Marasmitruncus massiliensis' gen. nov., sp. nov., 'Clostridium culturomicum' sp. nov., 'Blautia provencensis' sp. nov., 'Bacillus caccae' sp. nov. and 'Ornithinibacillus massiliensis' sp. nov., isolated from stool samples of undernourished African children ». In : *New Microbes and New Infections* 19, p. 38-42. ISSN : 2052-2975. DOI : 10.1016/j.nmni.2017.05.005.

Piel, W. H., M. J. Donoghue et M. J. Sanderson (2000). « TreeBASE : a database of phylogenetic knowledge ». In : *To the interoperable "Catalog of Life" with partners Species*, p. 41-47.

Pinevich, A. V. (2015). « Proposal to consistently apply the International Code of Nomenclature of Prokaryotes (ICNP) to names of the oxygenic photosynthetic bacteria (cyanobacteria), including those validly published under the International Code of Botanical Nomenclature (ICBN)/International Code of Nomenclature for algae, fungi and plants (ICN), and proposal to change Principle 2 of the ICNP ». In : *International Journal of Systematic and Evolutionary Microbiology,* 65.3, p. 1070-1074. ISSN : 1466-5026, DOI : 10.1099/ijs.0.000034.

Poirier, S., O. Rué, R. Peguilhan, G. Coeuret, M. Zagorec, M.-C. Champomier-Vergès, V. Loux et S. Chaillou (25 sept. 2018). « Deciphering intra-species bacterial diversity of meat and seafood spoilage microbiota using gyrB amplicon sequencing : A comparative analysis with 16S rDNA V3-V4 amplicon sequencing ». In : *PLoS ONE* 13.9. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0204629.

Prévot, A.-R. (juil. 1938). « Etudes de systématiques bactérienne IV Critique de la conception actuelle du genre Clostridium ». In : *Annales de l'Institut Pasteur : journal de microbiologie* 61, p. 72.

Price, M. N., P. S. Dehal et A. P. Arkin (10 mar. 2010). « FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments ». In : *PLOS ONE* 5.3, e9490. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0009490.

Pritchard, L., R. H. Glover, S. Humphris, J. G. Elphinstone et I. K. Toth (17 déc. 2015). « Genomics and taxonomy in diagnostics for food security : soft-rotting enterobacterial plant pathogens ». In : *Analytical Methods* 8.1, p. 12-24. ISSN : 1759-9679. DOI : 10.1039/C5AY02550H.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies et F. O. Glöckner (1er jan. 2013). « The SILVA ribosomal RNA gene database project : improved data processing and web-based tools ». In : *Nucleic Acids Research* 41 (D1), p. D590-D596. ISSN : 0305-1048. DOI : 10.1093/nar/gks1219.

Queiroz, K. de (3 mai 2005). « Ernst Mayr and the modern concept of species ». In : *Proceedings of the National Academy of Sciences of the United States of America* 102 (Suppl 1), p. 6600-6607. ISSN : 0027-8424. DOI : 10.1073/pnas.0502030102.

Rainey, F. A. (2009). « Family VIII. Ruminococcaceae fam. nov ». In : *Bergey's manual of systematic bacteriology,* 3, p. 1016.

Rainey, F. A. et P. H. Janssen (1er juin 1995). « Phylogenetic analysis by 16S ribosomal DNA sequence comparison reveals two unrelated groups of

species within the genus Ruminococcus ». In : *FEMS Microbiology Letters* 129.1, p. 69-73. ISSN : 0378-1097. DOI : 10.1111/j.1574-6968.1995.tb07559.x.

RAJILIĆ-STOJANOVIĆ, M. et W. M. de VOS (sept. 2014). « The first 1000 cultured species of the human gastrointestinal microbiota ». In : *Fems Microbiology Reviews* 38.5, p. 996-1047. ISSN : 0168-6445. DOI : 10.1111/1574-6976.12075.

READ, T., L. FORTUN-LAMOTHE, G. PASCAL, M. LE BOULCH, L. CAUQUIL, B. GABINAUD, C. BANNELIER, E. BALMISSE, N. DESTOMBES, O. BOUCHEZ, T. GIDENNE et S. COMBES (2019). « Diversity and Co-occurrence Pattern Analysis of Cecal Microbiota Establishment at the Onset of Solid Feeding in Young Rabbits ». In : *Frontiers in Microbiology* 10. ISSN : 1664-302X. DOI : 10.3389/fmicb.2019.00973.

REES, J. et K. CRANSTON (22 mai 2017). « Automated assembly of a reference taxonomy for phylogenetic data synthesis ». In : *Biodiversity Data Journal* 5, e12581. ISSN : 1314-2828. DOI : 10.3897/BDJ.5.e12581.

REISER, L., T. Z. BERARDINI, D. LI, R. MULLER, E. M. STRAIT, Q. LI, Y. MEZHERITSKY, A. VETUSHKO et E. HUALA (1er jan. 2016). « Sustainable funding for biocuration : The Arabidopsis Information Resource (TAIR) as a case study of a subscription-based funding model ». In : *Database* 2016. DOI : 10.1093/database/baw018.

RICHTER, M. et R. ROSSELLÓ-MÓRA (10 nov. 2009). « Shifting the genomic gold standard for the prokaryotic species definition ». In : *Proceedings of the National Academy of Sciences of the United States of America* 106.45, p. 19126-19131. ISSN : 0027-8424. DOI : 10.1073/pnas.0906412106.

ROBERTSON, C. E., J. K. HARRIS, J. R. SPEAR et N. R. PACE (1er déc. 2005). « Phylogenetic diversity and ecology of environmental Archaea ». In : *Current Opinion in Microbiology*. Growth development / edited by John N Reeve and Ruth A Schmitz 8.6, p. 638-642. ISSN : 1369-5274. DOI : 10.1016/j.mib.2005.10.003.

RODRIGUEZ-R, L. M. et K. T. KONSTANTINIDIS (2014). « Bypassing cultivation to identify bacterial species ». In : *Microbe* 9.3, p. 111-8.

ROGERS, D. J. et T. T. TANIMOTO (21 oct. 1960). « A Computer Program for Classifying Plants ». In : *Science* 132.3434, p. 1115-1118. ISSN : 0036-8075, 1095-9203. DOI : 10.1126/science.132.3434.1115.

Rognes, T., T. Flouri, B. Nichols, C. Quince et F. Mahé (18 oct. 2016). « VSEARCH : a versatile open source tool for metagenomics ». In : *PeerJ* 4, e2584. issn : 2167-8359. doi : 10.7717/peerj.2584.

Romero, P., J. Wagg, M. L. Green, D. Kaiser, M. Krummenacker et P. D. Karp (2005). « Computational prediction of human metabolic pathways from the complete human genome ». In : *Genome Biology* 6.1, R2. issn : 1465-6906. doi : 10.1186/gb-2004-6-1-r2.

Rosenberg, E., E. F. DeLong, E. Stackebrandt, S. Lory et F. Thompson, éd. (2013). *The Prokaryotes : Prokaryotic Biology and Symbiotic Associations*. 4e éd. The Prokaryotes. Berlin Heidelberg : Springer-Verlag. isbn : 978-3-642-30193-3.

Rosselló-Mora, R. et R. Amann (jan. 2001). « The species concept for prokaryotes ». In : *FEMS microbiology reviews* 25.1, p. 39-67. issn : 0168-6445. doi : 10.1111/j.1574-6976.2001.tb00571.x.

Rosselló-Móra, R. et R. Amann (1er juin 2015). « Past and future species definitions for Bacteria and Archaea ». In : *Systematic and Applied Microbiology*. Taxonomy in the age of genomics 38.4, p. 209-216. issn : 0723-2020. doi : 10.1016/j.syapm.2015.02.001.

Saitou, N. et M. Nei (1er juil. 1987). « The neighbor-joining method : a new method for reconstructing phylogenetic trees. » In : *Molecular Biology and Evolution* 4.4, p. 406-425. issn : 0737-4038. doi : 10.1093/oxfordjournals.molbev.a040454.

Sajed, T., A. Marcu, M. Ramirez, A. Pon, A. C. Guo, C. Knox, M. Wilson, J. R. Grant, Y. Djoumbou et D. S. Wishart (4 jan. 2016). « ECMDB 2.0 : A richer resource for understanding the biochemistry of E. coli ». In : *Nucleic Acids Research* 44 (D1), p. D495-501. issn : 1362-4962. doi : 10.1093/nar/gkv1060.

Sanderson, M. J., A. Purvis et C. Henze (1er mar. 1998). « Phylogenetic supertrees : Assembling the trees of life ». In : *Trends in Ecology & Evolution* 13.3, p. 105-109. issn : 0169-5347. doi : 10.1016/S0169-5347(97)01242-1.

Sarkar, S. F. et D. S. Guttman (avr. 2004). « Evolution of the Core Genome of Pseudomonas syringae, a Highly Clonal, Endemic Plant Pathogen ». In : *Applied and Environmental Microbiology* 70.4, p. 1999-2012. issn : 0099-2240. doi : 10.1128/AEM.70.4.1999-2012.2004.

Sasi Jyothsna, T. S., L. Tushar, C. Sasikala et C. V. Ramana (2016). « Paraclostridium benzoelyticum gen. nov., sp. nov., isolated from marine

sediment and reclassification of Clostridium bifermentans as Paraclostridium bifermentans comb. nov. Proposal of a new genus Paeniclostridium gen. nov. to accommodate Clostridium sordellii and Clostridium ghonii ». In : *International Journal of Systematic and Evolutionary Microbiology* 66.3, p. 1268-1274. DOI : 10.1099/ijsem.0.000874.

Sayers, E. W., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, I. Mizrachi, J. Ostell, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, E. Yaschenko et J. Ye (jan. 2009). « Database resources of the National Center for Biotechnology Information ». In : *Nucleic Acids Research* 37 (Database issue), p. D5-15. ISSN : 1362-4962. DOI : 10.1093/nar/gkn741.

Schildkraut, C. L., J. Marmur et P. Doty (1er juin 1962). « Determination of the base composition of deoxyribonucleic acid from its buoyant density in CsCl ». In : *Journal of Molecular Biology* 4.6, p. 430-443. ISSN : 0022-2836. DOI : 10.1016/S0022-2836(62)80100-4.

Schiraldi, C., M. Giuliano et M. De Rosa (sept. 2002). « Perspectives on biotechnological applications of archaea ». In : *Archaea* 1.2, p. 75-86. ISSN : 1472-3646.

Schleifer, K. H. et O. Kandler (déc. 1972). « Peptidoglycan types of bacterial cell walls and their taxonomic implications. » In : *Bacteriological Reviews* 36.4, p. 407-477. ISSN : 0005-3678.

Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn et C. F. Weber (2 oct. 2009). « Introducing mothur : Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities ». In : *Applied and Environmental Microbiology* 75.23, p. 7537-7541. ISSN : 0099-2240. DOI : 10.1128/AEM.01541-09.

Schnoes, A. M., S. D. Brown, I. Dodevski et P. C. Babbitt (2009). « Annotation Error in Public Databases : Misannotation of Molecular Function in Enzyme Superfamilies ». In : *PLOS Computational Biology* 5.12, e1000605. ISSN : 1553-7358. DOI : 10.1371/journal.pcbi.1000605.

Schomburg, I., L. Jeske, M. Ulbrich, S. Placzek, A. Chang et D. Schomburg (10 nov. 2017). « The BRENDA enzyme information system–From a database to an expert system ». In : *Journal of Biotechnology*. Bioinformatics Solutions for Big Data Analysis in Life Sciences presented by the German Network for Bioinformatics Infrastructure 261, p. 194-206. issn : 0168-1656. doi : 10.1016/j.jbiotec.2017.04.020.

Sczyrba, A., P. Hofmann, P. Belmann, D. Koslicki, S. Janssen, J. Dröge, I. Gregor, S. Majda, J. Fiedler, E. Dahms, A. Bremges, A. Fritz, R. Garrido-Oter, T. S. Jørgensen, N. Shapiro, P. D. Blood, A. Gurevich, Y. Bai, D. Turaev, M. Z. DeMaere, R. Chikhi, N. Nagarajan, C. Quince, F. Meyer, M. Balvočiūtė, L. H. Hansen, S. J. Sørensen, B. K. H. Chia, B. Denis, J. L. Froula, Z. Wang, R. Egan, D. Don Kang, J. J. Cook, C. Deltel, M. Beckstette, C. Lemaitre, P. Peterlongo, G. Rizk, D. Lavenier, Y.-W. Wu, S. W. Singer, C. Jain, M. Strous, H. Klingenberg, P. Meinicke, M. D. Barton, T. Lingner, H.-H. Lin, Y.-C. Liao, G. G. Z. Silva, D. A. Cuevas, R. A. Edwards, S. Saha, V. C. Piro, B. Y. Renard, M. Pop, H.-P. Klenk, M. Göker, N. C. Kyrpides, T. Woyke, J. A. Vorholt, P. Schulze-Lefert, E. M. Rubin, A. E. Darling, T. Rattei et A. C. McHardy (nov. 2017). « Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software ». In : *Nature Methods* 14.11, p. 1063-1071. issn : 1548-7105. doi : 10.1038/nmeth.4458.

Sebaihia, M., B. W. Wren, P. Mullany, N. F. Fairweather, N. Minton, R. Stabler, N. R. Thomson, A. P. Roberts, A. M. Cerdeño-Tárraga, H. Wang, M. T. Holden, A. Wright, C. Churcher, M. A. Quail, S. Baker, N. Bason, K. Brooks, T. Chillingworth, A. Cronin, P. Davis, L. Dowd, A. Fraser, T. Feltwell, Z. Hance, S. Holroyd, K. Jagels, S. Moule, K. Mungall, C. Price, E. Rabbinowitsch, S. Sharp, M. Simmonds, K. Stevens, L. Unwin, S. Whithead, B. Dupuy, G. Dougan, B. Barrell et J. Parkhill (juil. 2006). « The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome ». In : *Nature Genetics* 38.7, p. 779-786. issn : 1546-1718. doi : 10.1038/ng1830.

Seemann, T. (15 juil. 2014). « Prokka : rapid prokaryotic genome annotation ». In : *Bioinformatics* 30.14, p. 2068-2069. issn : 1367-4803. doi : 10.1093/bioinformatics/btu153.

Segata, N., D. Börnigen, X. C. Morgan et C. Huttenhower (2013). « PhyloPhlAn is a new method for improved phylogenetic and taxonomic

placement of microbes ». In : *Nature communications* 4, p. 2304. ISSN : 2041-1723. DOI : 10.1038/ncomms3304.

Segata, N. et C. Huttenhower (12 sept. 2011). « Toward an Efficient Method of Identifying Core Genes for Evolutionary and Functional Microbial Phylogenies ». In : *PLOS ONE* 6.9, e24704. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0024704.

Severi, E., D. W. Hood et G. H. Thomas (2007). « Sialic acid utilization by bacterial pathogens ». In : *Microbiology,* 153.9, p. 2817-2822. ISSN : 1350-0872, DOI : 10.1099/mic.0.2007/009480-0.

Shin, N.-R., W. Kang, E. J. Tak, D.-W. Hyun, P. S. Kim, H. S. Kim, J.-Y. Lee, H. Sung, T. W. Whon et J.-W. Bae (2018). « Blautia hominis sp. nov., isolated from human faeces ». In : *International Journal of Systematic and Evolutionary Microbiology* 68.4, p. 1059-1064. DOI : 10.1099/ijsem.0.002623.

Sijpesteijn, A. K. (1949). « Cellulose-decomposing bacteria from the rumen of cattle ». In : *Antonie van Leeuwenhoek* 15.1, p. 49-52.

Simpson, G. G. (1961). *Principles of Animal Taxonomy.* Columbia University Press. ISBN : 978-0-231-02427-3.

Sims, D., I. Sudbery, N. E. Ilott, A. Heger et C. P. Ponting (fév. 2014). « Sequencing depth and coverage : key considerations in genomic analyses ». In : *Nature Reviews. Genetics* 15.2, p. 121-132. ISSN : 1471-0064. DOI : 10.1038/nrg3642.

Skerman, V. B. D., V. McGowan et P. H. A. Sneath (1980). « Approved Lists of Bacterial Names ». In : *International Journal of Systematic and Evolutionary Microbiology* 30.1, p. 225-420. DOI : 10.1099/00207713-30-1-225.

Skerman, V. B. D., V. McGowan et P. H. A. Sneath, éd. (1989). *Approved Lists of Bacterial Names (Amended).* Washington (DC) : ASM Press. ISBN : 978-1-55581-014-6.

Slenter, D. N., M. Kutmon, K. Hanspers, A. Riutta, J. Windsor, N. Nunes, J. Mélius, E. Cirillo, S. L. Coort, D. Digles, F. Ehrhart, P. Giesbertz, M. Kalafati, M. Martens, R. Miller, K. Nishida, L. Rieswijk, A. Waagmeester, L. M. T. Eijssen, C. T. Evelo, A. R. Pico et E. L. Willighagen (4 jan. 2018). « WikiPathways : a multifaceted pathway database bridging metabolomics to other omics research ». In : *Nucleic Acids Research* 46 (D1), p. D661-D667. ISSN : 0305-1048. DOI : 10.1093/nar/gkx1064.

Smit, G., B. A. Smit et W. J. M. Engels (1er août 2005). « Flavour formation by lactic acid bacteria and biochemical flavour profiling of cheese products ». In : *FEMS Microbiology Reviews* 29.3, p. 591-610. issn : 0168-6445. doi : 10.1016/j.fmrre.2005.04.002.

Sneath, P. H. A. (2015). « Numerical Taxonomy ». In : *Bergey's Manual of Systematics of Archaea and Bacteria*. American Cancer Society, p. 1-5. isbn : 978-1-118-96060-8. doi : 10.1002/9781118960608.bm00018.

Sokal, R. R. (1963). « The Principles and Practice of Numerical Taxonomy ». In : *Taxon* 12.5, p. 190-199. issn : 0040-0262. doi : 10.2307/1217562.

Stackebrandt, E. et B. M. Goebel (1994). « Taxonomic Note : A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology ». In : *International Journal of Systematic and Evolutionary Microbiology* 44.4, p. 846-849. doi : 10.1099/00207713-44-4-846.

Stackebrandt, E. et J. Ebers (2006). « Taxonomic parameters revisited : tarnished gold standards ». In : *Microbiol. Today* 33, p. 152-155.

Stackebrandt, E., W. Frederiksen, G. M. Garrity, P. A. D. Grimont, P. Kämpfer, M. C. J. Maiden, X. Nesme, R. Rosselló-Mora, J. Swings, H. G. Trüper, L. Vauterin, A. C. Ward et W. B. Whitman (2002). « Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. » In : *International Journal of Systematic and Evolutionary Microbiology* 52.3, p. 1043-1047. doi : 10.1099/00207713-52-3-1043.

Stamatakis, A. (1er mai 2014). « RAxML version 8 : a tool for phylogenetic analysis and post-analysis of large phylogenies ». In : *Bioinformatics* 30.9, p. 1312-1313. issn : 1367-4803. doi : 10.1093/bioinformatics/btu033.

Stanier, R. Y. et C. B. van Niel (1er mar. 1962). « The concept of a bacterium ». In : *Archiv für Mikrobiologie* 42.1, p. 17-35. issn : 1432-072X. doi : 10.1007/BF00425185.

Stecher, B., L. Maier et W.-D. Hardt (avr. 2013). « 'Blooming' in the gut : how dysbiosis might contribute to pathogen evolution ». In : *Nature Reviews Microbiology* 11.4, p. 277-284. issn : 1740-1534. doi : 10.1038/nrmicro2989.

Stetter, K. O. (1er mai 1996). « Hyperthermophilic procaryotes ». In : *FEMS Microbiology Reviews* 18.2, p. 149-158. issn : 0168-6445. doi : 10.1111/j.1574-6976.1996.tb00233.x.

Stewart, E. J. (août 2012). « Growing Unculturable Bacteria ». In : *Journal of Bacteriology* 194.16, p. 4151-4160. issn : 0021-9193. doi : 10.1128/JB.00345-12.

Stoddard, S. F., B. J. Smith, R. Hein, B. R. Roller et T. M. Schmidt (28 jan. 2015). « rrnDB : improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development ». In : *Nucleic Acids Research* 43 (Database issue), p. D593-D598. issn : 0305-1048. doi : 10.1093/nar/gku1201.

Stubbendieck, R. M., C. Vargas-Bautista et P. D. Straight (8 août 2016). « Bacterial Communities : Interactions to Scale ». In : *Frontiers in Microbiology* 7. issn : 1664-302X. doi : 10.3389/fmicb.2016.01234.

Sueoka, N. (août 1961). « Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein ». In : *Proceedings of the National Academy of Sciences of the United States of America* 47.8, p. 1141-1149. issn : 0027-8424.

Suzek, B. E., Y. Wang, H. Huang, P. B. McGarvey et C. H. Wu (15 mar. 2015). « UniRef clusters : a comprehensive and scalable alternative for improving sequence similarity searches ». In : *Bioinformatics* 31.6, p. 926-932. issn : 1367-4803. doi : 10.1093/bioinformatics/btu739.

Tardieu, C. (1er déc. 2011). « La bonne orthographe du mot taxinomie. Un concept important dont l'orthographe est malmenée ». In : *PALEO. Revue d'archéologie préhistorique* 22, p. 331-334. issn : 1145-3370.

Tatusova, T., S. Ciufo, B. Fedorov, K. O'Neill et I. Tolstoy (jan. 2014). « RefSeq microbial genomes database : new representation and annotation strategy ». In : *Nucleic Acids Research* 42 (Database issue), p. D553-559. issn : 1362-4962. doi : 10.1093/nar/gkt1274.

Tatusova, T., M. DiCuccio, A. Badretdin, V. Chetvernin, E. P. Nawrocki, L. Zaslavsky, A. Lomsadze, K. D. Pruitt, M. Borodovsky et J. Ostell (19 août 2016). « NCBI prokaryotic genome annotation pipeline ». In : *Nucleic Acids Research* 44.14, p. 6614-6624. issn : 0305-1048. doi : 10.1093/nar/gkw569.

Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V. Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. DeBoy, T. M. Davidsen, M. Mora, M. Scarselli, I. M. y. Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C. Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C. Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D. Radune, G. Dimitrov, K. Watkins,

K. J. B. O'Connor, S. Smith, T. R. Utterback, O. White, C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R. Rappuoli et C. M. Fraser (27 sept. 2005). « Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae : Implications for the microbial "pan-genome" ». In : *Proceedings of the National Academy of Sciences of the United States of America* 102.39, p. 13950. doi : 10.1073/pnas.0506758102.

The Human Microbiome Project Consortium et al. (juin 2012). « Structure, function and diversity of the healthy human microbiome ». In : *Nature* 486.7402, p. 207-214. issn : 1476-4687. doi : 10.1038/nature11234.

Thomas, C. M. et K. M. Nielsen (sept. 2005). « Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria ». In : *Nature Reviews Microbiology* 3.9, p. 711. issn : 1740-1534. doi : 10.1038/nrmicro1234.

Thompson, J. R., S. Pacocha, C. Pharino, V. Klepac-Ceraj, D. E. Hunt, J. Benoit, R. Sarma-Rupavtarm, D. L. Distel et M. F. Polz (25 fév. 2005). « Genotypic Diversity Within a Natural Coastal Bacterioplankton Population ». In : *Science* 307.5713, p. 1311-1313. issn : 0036-8075, 1095-9203. doi : 10.1126/science.1106028.

Tindall, B. J., R. Rosselló-Móra, H.-J. Busse, W. Ludwig et P. Kämpfer (2010). « Notes on the characterization of prokaryote strains for taxonomic purposes ». In : *International Journal of Systematic and Evolutionary Microbiology* 60.1, p. 249-266. doi : 10.1099/ijs.0.016949-0.

Togo, A. H., A. Diop, F. Bittar, M. Maraninchi, R. Valero, N. Armstrong, G. Dubourg, N. Labas, M. Richez, J. Delerce, A. Levasseur, P.-E. Fournier, D. Raoult et M. Million (1er nov. 2018). « Description of Mediterraneibacter massiliensis, gen. nov., sp. nov., a new genus isolated from the gut microbiota of an obese patient and reclassification of Ruminococcus faecis, Ruminococcus lactaris, Ruminococcus torques, Ruminococcus gnavus and Clostridium glycyrrhizinilyticum as Mediterraneibacter faecis comb. nov., Mediterraneibacter lactaris comb. nov., Mediterraneibacter torques comb. nov., Mediterraneibacter gnavus comb. nov. and Mediterraneibacter glycyrrhizinilyticus comb. nov. » In : *Antonie van Leeuwenhoek* 111.11, p. 2107-2128. issn : 1572-9699. doi : 10.1007/s10482-018-1104-y.

Torsvik, V., L. Øvreås et T. F. Thingstad (10 mai 2002). « Prokaryotic Diversity–Magnitude, Dynamics, and Controlling Factors ». In : *Science* 296.5570, p. 1064-1066. issn : 0036-8075, 1095-9203. doi : 10.1126/science.1071698.

Traore, S., E. Azhar, M. Yasir, F. Bibi, P.-E. Fournier, A. Jiman-Fatani, J. Delerce, F. Cadoret, J.-C. Lagier et D. Raoult (3 juin 2017). « Description of 'Blautia phocaeensis' sp. nov. and 'Lachnoclostridium edouardi' sp. nov., isolated from healthy fresh stools of Saudi Arabia Bedouins by culturomics ». In : *New Microbes and New Infections* 19, p. 129-131. issn : 2052-2975. doi : 10.1016/j.nmni.2017.05.017.

Travers, M., S. M. Paley, J. Shrager, T. A. Holland et P. D. Karp (1er jan. 2013). « Groups : knowledge spreadsheets for symbolic biocomputing ». In : *Database* 2013. doi : 10.1093/database/bat061.

Treangen, T. J. et E. P. C. Rocha (27 jan. 2011). « Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes ». In : *PLoS Genetics* 7.1. issn : 1553-7390. doi : 10.1371/journal.pgen.1001284.

Truong, D. T., E. A. Franzosa, T. L. Tickle, M. Scholz, G. Weingart, E. Pasolli, A. Tett, C. Huttenhower et N. Segata (oct. 2015). « MetaPhlAn2 for enhanced metagenomic taxonomic profiling ». In : *Nature Methods* 12.10, p. 902-903. issn : 1548-7105. doi : 10.1038/nmeth.3589.

Turland, N., J. Wiersema, F. Barrie, W. Greuter, D. Hawksworth, P. Herendeen, S. Knapp, W.-H. Kusber, D.-Z. Li, K. Marhold, T. May, J. McNeill, A. Monro, J. Prado, M. Price et G. Smith, éd. (26 juin 2018). *International Code of Nomenclature for algae, fungi, and plants*. T. 159. Regnum Vegetabile. Koeltz Botanical Books. isbn : 978-3-946583-16-5. doi : 10.12705/Code.2018.

Turnbaugh, P. J., R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight et J. I. Gordon (17 oct. 2007). « The Human Microbiome Project ». In : *Nature* 449, p. 804-810. issn : 1476-4687. doi : 10.1038/nature06244.

Tyson, G. W., J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar et J. F. Banfield (mar. 2004). « Community structure and metabolism through reconstruction of microbial genomes from the environment ». In : *Nature* 428.6978, p. 37-43. issn : 1476-4687. doi : 10.1038/nature02340.

Valas, R. E. et P. E. Bourne (25 août 2009). « Structural analysis of polarizing indels : an emerging consensus on the root of the tree of life ». In : *Biology Direct* 4, p. 30. issn : 1745-6150. doi : 10.1186/1745-6150-4-30.

Vallenet, D., A. Calteau, S. Cruveiller, M. Gachet, A. Lajus, A. Josso, J. Mercier, A. Renaux, J. Rollin, Z. Rouy, D. Roche, C. Scarpelli et C. Médigue (4 jan. 2017). « MicroScope in 2017 : an expanding and evolving integrated resource for community expertise of microbial genomes ». In : *Nucleic Acids Research* 45 (D1), p. D517-D528. issn : 0305-1048. doi : 10.1093/nar/gkw1101.

Vallenet, D., S. Engelen, D. Mornico, S. Cruveiller, L. Fleury, A. Lajus, Z. Rouy, D. Roche, G. Salvignol, C. Scarpelli et C. Médigue (1er jan. 2009). « MicroScope : a platform for microbial genome annotation and comparative genomics ». In : *Database* 2009. doi : 10.1093/database/bap021.

Vallenet, D., L. Labarre, Z. Rouy, V. Barbe, S. Bocs, S. Cruveiller, A. Lajus, G. Pascal, C. Scarpelli et C. Médigue (1er jan. 2006). « MaGe : a microbial genome annotation system supported by synteny results ». In : *Nucleic Acids Research* 34.1, p. 53-65. issn : 0305-1048. doi : 10.1093/nar/gkj406.

Vandamme, P., B. Pot, M. Gillis, P. de Vos, K. Kersters et J. Swings (juin 1996). « Polyphasic taxonomy, a consensus approach to bacterial systematics. » In : *Microbiological Reviews* 60.2, p. 407-438. issn : 0146-0749.

Vartoukian, S. R., R. M. Palmer et W. G. Wade (1er août 2010). « Strategies for culture of 'unculturable' bacteria ». In : *FEMS Microbiology Letters* 309.1, p. 1-7. issn : 0378-1097. doi : 10.1111/j.1574-6968.2010.02000.x.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.-H. Rogers et H. O. Smith (2 avr. 2004). « Environmental Genome Shotgun Sequencing of the Sargasso Sea ». In : *Science* 304.5667, p. 66-74. issn : 0036-8075, 1095-9203. doi : 10.1126/science.1093857.

Větrovský, T. et P. Baldrian (27 fév. 2013). « The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses ». In : *PLoS ONE* 8.2. issn : 1932-6203. doi : 10.1371/journal.pone.0057923.

Vincent, A. T., N. Derome, B. Boyle, A. I. Culley et S. J. Charette (1er juil. 2017). « Next-generation sequencing (NGS) in the microbiolo-

gical world : How to make the most of your money ». In : *Journal of Microbiological Methods*. What's next in microbiology methods ? Emerging methods 138, p. 60-71. ISSN : 0167-7012. DOI : 10.1016/j.mimet.2016.02.016.

Vos, P., G. Garrity, D. Jones, N. R. Krieg, W. Ludwig, F. A. Rainey, K.-H. Schleifer et W. Whitman, éd. (2009). *Bergey's Manual of Systematic Bacteriology : Volume 3 : The Firmicutes*. 2e éd. Bergey's Manual of Systematic Bacteriology. New York : Springer-Verlag. ISBN : 978-0-387-95041-9.

Wadi, L., M. Meyer, J. Weiser, L. D. Stein et J. Reimand (sept. 2016). « Impact of outdated gene annotations on pathway enrichment analysis ». In : *Nature Methods* 13.9, p. 705-706. ISSN : 1548-7091. DOI : 10.1038/nmeth.3963.

Wagner, M. et A. Loy (1er juin 2002). « Bacterial community composition and function in sewage treatment systems ». In : *Current Opinion in Biotechnology* 13.3, p. 218-227. ISSN : 0958-1669. DOI : 10.1016/S0958-1669(02)00315-4.

Waite, D. W., I. Vanwonterghem, C. Rinke, D. H. Parks, Y. Zhang, K. Takai, S. M. Sievert, J. Simon, B. J. Campbell, T. E. Hanson, T. Woyke, M. G. Klotz et P. Hugenholtz (2017). « Comparative Genomic Analysis of the Class Epsilonproteobacteriaand Proposed Reclassification to Epsilonbacteraeota (phyl. nov.) » In : *Frontiers in Microbiology* 8, p. 682. ISSN : 1664-302X. DOI : 10.3389/fmicb.2017.00682.

Wang, Q., G. M. Garrity, J. M. Tiedje et J. R. Cole (août 2007). « Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy ». In : *Applied and Environmental Microbiology* 73.16, p. 5261-5267. ISSN : 0099-2240. DOI : 10.1128/AEM.00062-07.

Wattam, A. R., J. J. Davis, R. Assaf, S. Boisvert, T. Brettin, C. Bun, N. Conrad, E. M. Dietrich, T. Disz, J. L. Gabbard, S. Gerdes, C. S. Henry, R. W. Kenyon, D. Machi, C. Mao, E. K. Nordberg, G. J. Olsen, D. E. Murphy-Olson, R. Olson, R. Overbeek, B. Parrello, G. D. Pusch, M. Shukla, V. Vonstein, A. Warren, F. Xia, H. Yoo et R. L. Stevens (4 jan. 2017). « Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center ». In : *Nucleic Acids Research* 45 (Database issue), p. D535-D542. ISSN : 0305-1048. DOI : 10.1093/nar/gkw1017.

Wayne, L. G., D. J. Brenner, R. R. Colwell, P. A. D. Grimont, O. Kandler, M. I. Krichevsky, L. H. Moore, W. E. C. Moore, R. G. E.

Murray, E. Stackebrandt, M. P. Starr et H. G. Truper (1987). « Report of the Ad Hoc Committee on Reconciliation of Approaches to Bacterial Systematics ». In : *International Journal of Systematic and Evolutionary Microbiology* 37.4, p. 463-464. doi : 10.1099/00207713-37-4-463.

Wegmann, U., P. Louis, A. Goesmann, B. Henrissat, S. H. Duncan et H. J. Flint (sept. 2014). « Complete genome of a new Firmicutes species belonging to the dominant human colonic microbiota ('Ruminococcus bicirculans') reveals two chromosomes and a selective capacity to utilize plant glucans ». In : *Environmental Microbiology* 16.9, p. 2879-2890. issn : 1462-2920. doi : 10.1111/1462-2920.12217.

Weisburg, W. G., S. M. Barns, D. A. Pelletier et D. J. Lane (jan. 1991). « 16S ribosomal DNA amplification for phylogenetic study. » In : *Journal of Bacteriology* 173.2, p. 697-703. issn : 0021-9193.

Wemheuer, F., J. A. Taylor, R. Daniel, E. Johnston, P. Meinicke, T. Thomas et B. Wemheuer (9 déc. 2018). « Tax4Fun2 : a R-based tool for the rapid prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene marker gene sequences ». In : *bioRxiv*, p. 490037. doi : 10.1101/490037.

Wetterstrand, K. (2019). *DNA Sequencing Costs : Data from the NHGRI Genome Sequencing Program*. Genome.gov. url : https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data (visité le 15/09/2019).

Whitcomb, R. F. (2007). « Evolution and devolution of minimal standards for descriptions of species of the class Mollicutes : analysis of two Spiroplasma descriptions ». In : *International Journal of Systematic and Evolutionary Microbiology,* 57.2, p. 201-206. issn : 1466-5026, doi : 10.1099/ijs.0.64545-0.

Whitman, W. B. (avr. 2009). « The Modern Concept of the Procaryote ». In : *Journal of Bacteriology* 191.7, p. 2000-2005. issn : 0021-9193. doi : 10.1128/JB.00962-08.

Whitman, W. B., D. C. Coleman et W. J. Wiebe (9 juin 1998). « Prokaryotes : The unseen majority ». In : *Proceedings of the National Academy of Sciences* 95.12, p. 6578-6583. issn : 0027-8424, 1091-6490.

Whitman, W. B., I. C. Sutcliffe et R. Rossello-Mora (juil. 2019). « Proposal for changes in the International Code of Nomenclature of Prokaryotes : granting priority to Candidatus names ». In : *International*

Journal of Systematic and Evolutionary Microbiology 69.7, p. 2174-2175. ISSN : 1466-5034. DOI : 10.1099/ijsem.0.003419.

WHITMAN, W., M. GOODFELLOW, P. KÄMPFER, H.-J. BUSSE, M. TRUJILLO, W. LUDWIG, K.-i. SUZUKI et A. PARTE, éd. (2012). *Bergey's Manual of Systematic Bacteriology : Volume 5 : The Actinobacteria*. 2e éd. Bergey's Manual of Systematic Bacteriology. New York : Springer-Verlag. ISBN : 978-0-387-95043-3.

WILKINS, J. S. (5 fév. 2011). *What is systematics and what is taxonomy ?* Evolving Thoughts. URL : https://evolvingthoughts.net/2011/02/05/what-is-systematics-and-what-is-taxonomy/'.

WILLEY, J. M., L. SHERWOOD, C. J. WOOLVERTON et L. M. PRESCOTT (2017). *Prescott's microbiology*. OCLC : 934479924. ISBN : 978-1-259-28159-4.

WILSON, R. A. (1999). *Species : New Interdisciplinary Essays*. MIT Press.

WISHART, D. S., Y. D. FEUNANG, A. MARCU, A. C. GUO, K. LIANG, R. VÁZQUEZ-FRESNO, T. SAJED, D. JOHNSON, C. LI, N. KARU, Z. SAYEEDA, E. LO, N. ASSEMPOUR, M. BERJANSKII, S. SINGHAL, D. ARNDT, Y. LIANG, H. BADRAN, J. GRANT, A. SERRA-CAYUELA, Y. LIU, R. MANDAL, V. NEVEU, A. PON, C. KNOX, M. WILSON, C. MANACH et A. SCALBERT (4 jan. 2018). « HMDB 4.0 : the human metabolome database for 2018 ». In : *Nucleic Acids Research* 46 (Database issue), p. D608-D617. ISSN : 0305-1048. DOI : 10.1093/nar/gkx1089.

WOESE, C. R. (juin 1987). « Bacterial evolution. » In : *Microbiological Reviews* 51.2, p. 221-271. ISSN : 0146-0749.

WOESE, C. R., O. KANDLER et M. L. WHEELIS (juin 1990). « Towards a natural system of organisms : proposal for the domains Archaea, Bacteria, and Eucarya. » In : *Proceedings of the National Academy of Sciences of the United States of America* 87.12, p. 4576-4579. ISSN : 0027-8424.

WOESE, C. R. et G. E. FOX (1er nov. 1977). « Phylogenetic structure of the prokaryotic domain : The primary kingdoms ». In : *Proceedings of the National Academy of Sciences* 74.11, p. 5088-5090. ISSN : 0027-8424, 1091-6490. DOI : 10.1073/pnas.74.11.5088.

WOOD, D. E. et S. L. SALZBERG (2014). « Kraken : ultrafast metagenomic sequence classification using exact alignments ». In : *Genome Biology* 15.3, R46. ISSN : 1465-6906. DOI : 10.1186/gb-2014-15-3-r46.

Yang, Z. et B. Rannala (mai 2012). « Molecular phylogenetics : principles and practice ». In : *Nature Reviews Genetics* 13.5, p. 303-314. issn : 1471-0064. doi : 10.1038/nrg3186.

Yarza, P., W. Ludwig, J. Euzéby, R. Amann, K.-H. Schleifer, F. O. Glöckner et R. Rosselló-Móra (1er oct. 2010). « Update of the All-Species Living Tree Project based on 16S and 23S rRNA sequence analyses ». In : *Systematic and Applied Microbiology* 33.6, p. 291-299. issn : 0723-2020. doi : 10.1016/j.syapm.2010.08.001.

Yarza, P., M. Richter, J. Peplies, J. Euzeby, R. Amann, K.-H. Schleifer, W. Ludwig, F. O. Glöckner et R. Rosselló-Móra (1er sept. 2008). « The All-Species Living Tree project : A 16S rRNA-based phylogenetic tree of all sequenced type strains ». In : *Systematic and Applied Microbiology* 31.4, p. 241-250. issn : 0723-2020. doi : 10.1016/j.syapm.2008.07.001.

Yarza, P., P. Yilmaz, E. Pruesse, F. O. Glöckner, W. Ludwig, K.-H. Schleifer, W. B. Whitman, J. Euzéby, R. Amann et R. Rosselló-Móra (sept. 2014). « Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences ». In : *Nature Reviews Microbiology* 12.9, p. 635-645. issn : 1740-1534. doi : 10.1038/nrmicro3330.

Ye, Y. et T. G. Doak (2009). « A Parsimony Approach to Biological Pathway Reconstruction/Inference for Genomes and Metagenomes ». In : *PLOS Computational Biology* 5.8, e1000465. issn : 1553-7358. doi : 10.1371/journal.pcbi.1000465.

Yilmaz, P., L. W. Parfrey, P. Yarza, J. Gerken, E. Pruesse, C. Quast, T. Schweer, J. Peplies, W. Ludwig et F. O. Glöckner (1er jan. 2014). « The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks ». In : *Nucleic Acids Research* 42 (Database issue), p. D643-D648. issn : 0305-1048. doi : 10.1093/nar/gkt1209.

Yoon, S.-H., S.-M. Ha, S. Kwon, J. Lim, Y. Kim, H. Seo et J. Chun (mai 2017a). « Introducing EzBioCloud : a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies ». In : *International Journal of Systematic and Evolutionary Microbiology* 67.5, p. 1613-1617. issn : 1466-5026. doi : 10.1099/ijsem.0.001755.

Yoon, S.-H., S.-M. Ha, J. Lim, S. Kwon et J. Chun (1er oct. 2017b). « A large-scale evaluation of algorithms to calculate average nucleotide identity ». In : *Antonie van Leeuwenhoek* 110.10, p. 1281-1286. issn : 1572-9699. doi : 10.1007/s10482-017-0844-4.

Zakhia, F. et P. de Lajudie (1er mar. 2006). « La taxonomie bactérienne moderne : revue des techniques — application à la caractérisation des bactéries nodulant les légumineuses (BNL) ». In : *Canadian Journal of Microbiology* 52.3, p. 169-181. ISSN : 0008-4166. DOI : 10.1139/w05-092.

Zepeda Mendoza, M. L., T. Sicheritz-Pontén et M. T. P. Gilbert (sept. 2015). « Environmental genes and genomes : understanding the differences and challenges in the approaches and software for their analyses ». In : *Briefings in Bioinformatics* 16.5, p. 745-758. ISSN : 1467-5463. DOI : 10.1093/bib/bbv001.

Zhang, X., B. Tu, L.-r. Dai, P. A. Lawson, Z.-z. Zheng, L.-Y. Liu, Y. Deng, H. Zhang et L. Cheng (2018). « Petroclostridium xylanilyticum gen. nov., sp. nov., a xylan-degrading bacterium isolated from an oilfield, and reclassification of clostridial cluster III members into four novel genera in a new Hungateiclostridiaceae fam. nov. » In : *International Journal of Systematic and Evolutionary Microbiology,* 68.10, p. 3197-3211. ISSN : 1466-5026, DOI : 10.1099/ijsem.0.002966.

Abstract

Prokaryotes are ubiquitous organisms living in communities, whose extreme metabolic diversity is correlated with their ubiquity. These characteristics have led Man to identify, name, classify and attempt to understand their role within communities, in order to shape these communities and, ultimately, their environment.

To contribute to a better understanding of the functional role of prokaryotes, we developed MACADAM : a database of metabolic pathways associated with a prokaryote-centric taxonomy. The aim is to provide the scientific community with open access to functional information data which has been selected for its genomic and annotation quality, which is interoperable and simply structured, thereby enabling updates to be made to the data gathered from data sources such as MetaCyc, MicroCyc and RefSeq by MACADAM. MACADAM meets these criteria. MACADAM includes PGDBs (Pathway/Genome DataBases) assembled from RefSeq genomes meeting the « complete genome » quality criteria, by using the Pathway Tools software made available by MetaCyc, a metabolic pathway database. In order to enrich the database and increase the quality of functional information in MACADAM, a collection of expert-curated PGDBs named MicroCyc was added. Its PGDBs are favoured over those of RefSeq. Functional information sourced from the literature contained in FAPROTAX and IJSEM phenotypic databases was also added. MACADAM contains 13 509 PGDBs (13 195 bacterial PGDBs and 314 archaeal PGDBs) and 1 260 unique metabolic pathways. Built using interoperable technologies (Python 3, SQLite), in a downloadable format and with open-source code, MACADAM can be integrated into tools requiring the pairing of functional and taxonomic information. To improve its visibility among the microbiology community, MACADAM is available online (`http://macadam.toulouse.inra.fr`). By using the taxonomy of the « NCBI Taxonomy » database, MACADAM makes it possible to link any taxon—ranging from phylum to species—to its functional information. Each metabolic pathway is associated with two completeness scores (a PS : Pathway Score and a PFS : Pathway Frequency Score). With each update, MACADAM integrates the new versions of RefSeq, NCBI Taxonomy and MicroCyc, allowing any corrections made to the taxonomy to be promptly amended and to add information on recently-submitted genomes.

Two examples of ways in which to use MACADAM, and a comparison with an inference approach based on metagenomic readings allowed for a discussion of the strengths and weaknesses (i) of MACADAM and (ii) of inference by a prior taxonomic identification approach. The identification of individuals within the prokaryotic community benefits greatly from advances in sequencing technology and the refinement of bioinformatics analysis pipelines. The analysis of readings from metagenomic sequencing leads to the reconstruction of putative genomes and metagenomic species. In this context, we examined the problem of correcting taxonomic assignments of metagenomic species, by using a phylogenetic tree reconstruction approach on the one hand, and by using an overall genome relatedness index (ANI) on the other hand. This work allowed us to clarify the positioning of nine groups of metagenomic species, and highlighted errors in reference genome affiliation in *Megasphaera* and *Blautia Obeum*. It also allowed us to confirm the reclassification of *Ruminococcus gauvreauii* into the genus *Blautia*. To limit errors and prevent their replication, it is important to ensure the quality of the information contained in the databases. In this context, the scientific community should have better knowledge of the rules of nomenclature and systematic methods. Further efforts should be made to advocate the merits of correcting database data. Finally, although metagenomics provides a better understanding of the microbial communities around us, an effort to cultivate organisms that are said to be uncultivable would increase the knowledge and diversity of prokaryotic organisms in databases. These efforts will have a direct impact on the quality of functional information and the coverage of MACADAM's prokaryotic diversity.

Keywords : Taxonomy, Prokaryote, Functional Inference, Metabolic Pathways, Database

Résumé

Les procaryotes sont des organismes ubiquitaires vivant en communauté et possédant une extrême diversité métabolique en lien avec leur omniprésence. Ces caractéristiques ont poussé l'Homme à les identifier, les nommer, les classer et comprendre leur rôle au sein des communautés afin de modeler ces communautés et *in fine* leur environnement.

Pour contribuer à la compréhension du rôle fonctionnel des procaryotes, nous avons développé MACADAM : une base de données de voies métaboliques associées à une taxonomie centrée sur les procaryotes. L'objectif était de mettre à disposition de la communauté scientifique des données d'informations fonctionnelles sélectionnées sur leur qualité (qualité des génomes, qualité des annotations), en accès libre, interopérables et avec une structure simple permettant des mises à jour afin de bénéficier des dernières versions des sources de données utilisées par MACADAM (MetaCyc, MicroCyc, RefSeq). MACADAM remplit ces critères. MACADAM regroupe les PGDBs (Pathway/Genome DataBase) construites à partir de génome RefSeq répondant aux critères de qualité « complete genome » en utilisant le logiciel Pathway Tools mis à disposition par la base de données de voies métaboliques MetaCyc. Afin d'enrichir la base et d'augmenter la qualité des informations fonctionnelles dans MACADAM, MicroCyc, une collection de PGDBs manuellement curées par des experts, a été ajoutée et préférée en cas de redondance vis-à-vis des PGDBs issues de RefSeq. Enfin, les informations fonctionnelles sourcées à partir de la littérature contenues dans FAPROTAX et IJSEM phenotypic databases sont ajoutées. MACADAM contient 13 509 PGDBs (13 195 PGDBs bactériennes et 314 PGDBs d'archées) et 1 260 voies métaboliques uniques. Construit à l'aide de technologies interopérables (Python 3, SQLite), sous un format téléchargeable et avec un code ouvert, MACADAM peut être intégré dans des outils qui nécessitent de lier une information taxonomique à une information fonctionnelle. Pour améliorer sa visibilité auprès de la communauté de microbiologistes, MACADAM est consultable en ligne (`http://macadam.toulouse.inra.fr`). Utilisant la taxonomie de la base de données « NCBI Taxonomy », MACADAM permet de relier un taxon allant du phylum à l'espèce à une information fonctionnelle. Chaque voie métabolique est associée à deux scores de complétude (PS : Pathway Score et PFS : Pathway Frequency Score). A chaque mise à jour, MACADAM intègre les nouvelles versions de RefSeq, du NCBI Taxonomy et de MicroCyc, permettant de suivre au plus près les corrections apportées à la taxonomie et d'inclure les informations disponibles pour les nouveaux génomes déposés.

Deux exemples d'utilisation de MACADAM et une comparaison avec une approche d'inférence à partir de lectures métagénomiques ont permis de discuter les points forts et les faiblesses (i) de MACADAM et (ii) de l'inférence par une approche d'identification taxonomique préalable. L'identification des individus au sein de la communauté procaryote bénéficie largement des avancées en technologie de séquençage et du raffinement des pipelines d'analyses bioinformatiques. L'analyse des lectures issues de séquençages métagénomiques aboutit à la reconstruction de génomes putatifs ou espèces métagénomiques. Dans ce cadre, nous nous sommes penchés sur la problématique de correction d'assignation taxonomique d'espèces métagénomiques en utilisant une approche par reconstruction d'un arbre phylogénétique d'une part et en utilisant un indice global de parenté génomique (ANI) d'autre part. Ce travail nous a permis de préciser le positionnement de neuf groupes d'espèces métagénomiques et mis en évidence des erreurs d'affiliation de génome de référence chez *Megasphaera* et *Blautia Obeum* et de confirmer le reclassement de *Ruminococcus gauvreauii* dans le genre *Blautia*. Pour limiter les erreurs et leur réplication il convient de veiller à la qualité de l'information contenue dans les bases de données. Dans ce cadre, la communauté scientifique devrait avoir une meilleure connaissance des règles de la nomenclature et des méthodes de systématique. Un intérêt accru devrait être porté pour valoriser les efforts de correction des données présentes dans les bases de données. Enfin, bien que la métagénomique permette de mieux comprendre les communautés microbiennes qui nous entourent, un effort de culture des organismes réputés incultivables permettrait d'accroître les connaissances et la diversité des organismes procaryotes dans les banques de données. Ces efforts se répercuteront directement sur la qualité des informations fonctionnelles et la couverture de la diversité des procaryotes de MACADAM.

Mots clés : Taxonomie, Procaryotes, Inférence Fonctionnelle, Voies Métaboliques, Base de données