



Génomique et métagénomique comparatives des bactéries

Thèse

Maxime Déraspe

Doctorat en médecine moléculaire
Philosophiæ doctor (Ph. D.)

Québec, Canada

Génomique et métagénomique comparatives des bactéries

Thèse

Maxime Déraspe

Sous la direction de:

Jacques Corbeil, directeur de recherche
François Laviolette, codirecteur de recherche

Résumé

Les domaines de la génomique et de la métagénomique ont apporté un support incommensurable à l'avancement de nos connaissances sur la génétique des bactéries. Les bactéries pathogènes sont maintenant séquencées et analysées pour identifier les facteurs causant leur virulence et/ou leur résistance aux antibiotiques ainsi que leur capacité à transmettre ces éléments génétiques qui sont d'un intérêt clinique. Les bactéries commensales, quant à elles, sont de plus en plus associées à la santé humaine et sont étudiées à l'aide de la métagénomique pour contrer les difficultés liées à leur culture étant donné leur grande diversité en matière de besoins métaboliques. Les nouvelles technologies de séquençages permettent donc de produire en masse ces séquences d'ADN à des fins de caractérisation et de comparaison dans le but d'élucider des questions souvent reliées à la santé humaine.

Les avancées en génomique et en métagénomique requièrent des logiciels bio-informatiques capables de gérer et de s'adapter à la quantité massive et croissante des données biologiques. Les deux premières hypothèses de ce doctorat concernaient le développement de méthodes efficaces et flexibles pour l'analyse de génomes et de métagénomes bactériens. Plusieurs méthodes d'analyses bio-informatiques ont été explorées et ont mené à l'implémentation de deux logiciels pour supporter les hypothèses de recherche : Ray Surveyor et kAamer.

La première hypothèse de recherche consistait à vérifier s'il était possible d'obtenir une comparaison de génomes, depuis leur simple contenu en k -mers de séquences d'ADN, avec des résultats analogues aux comparaisons génomiques standards comme le pourcentage moyen d'identités ou les arbres phylogénétiques, mais sans nécessiter d'alignements de séquences. Nous avons démontré avec le logiciel Ray Surveyor et plusieurs analyses de génomique et de métagénomique bactérienne, qu'il était possible d'obtenir de tels comparaisons à l'aide de séquences d'ADN découpées en k -mer. Dans l'étude qui présente les résultats de l'hypothèse de recherche, nous avons aussi estimé la propension génotypique de plusieurs espèces bactériennes à des phénotypes d'intérêt clinique à l'aide de bases de données de gènes spécialisées.

La deuxième hypothèse était de tester s'il était possible de développer un logiciel pour l'identification de séquences protéiques, basé sur des k -mers d'acides aminés, qui serait plus performant que les logiciels existants, spécifiquement pour l'identification de protéines avec un haut degré d'homologie. Les travaux menèrent à l'implémentation de kAamer, un logiciel permettant de

créer des bases de données de protéines où la recherche de séquence se fait par association exacte de k -mers tout en supportant l'alignement de séquences. KAAmer s'est avéré très efficace pour la recherche de séquences de protéines avec des performances surpassant même, dans la majorité des scénarios, les aligneurs de séquences les plus rapides. D'autres fonctionnalités intéressantes sont aussi offertes par kAAmer, tel que la possibilité d'héberger une base de données en tant que service de manière permanente.

Enfin, la troisième et dernière hypothèse de recherche visait à valider si les deux logiciels développés durant le projet de doctorat (Ray Surveyor et kAAmer) produiraient des résultats viables dans une analyse métagénomique du microbiote intestinal en lien avec l'obésité. Les profilages taxonomique et fonctionnel furent donc réalisés avec kAAmer et la comparaison *de novo* des métagénomomes investiguée avec Ray Surveyor. Les résultats obtenus se sont avérés significatifs et ont démontrés, entre autres, une tendance vers une abondance relative plus élevée pour le phylum *Bacteroidetes* et moins élevée pour les phyla *Firmicutes* et *Acinetobacteria* chez les sujets obèses. Une multitude de fonctions métaboliques se sont aussi avérées significativement différentes dans les conditions normales et d'obésités des métagénomomes, avec une mention particulière à celles reliées au métabolisme des acides gras à chaîne courte qui sont reconnues pour être associées à l'obésité.

Abstract

The fields of genomics and metagenomics have provided immeasurable support to the advancement of our knowledge of bacterial genetics. Pathogenic bacteria are now routinely sequenced and analyzed to identify the factors causing their virulence or antibiotic resistance as well as their ability to transmit genetic elements. Commensal bacteria are increasingly associated with human health and are being studied using metagenomics to counter the issues associated with their culture due to their wide range of metabolic needs. Next generation sequencing enabled us to mass-produce these DNA sequences for characterization and comparison purposes in order to elucidate questions related to human health.

Improvement in genomics and metagenomics studies required bio-informatics software that are able to manage and adapt to an increasing availability of biological sequences data. The first two hypotheses of this thesis include the development of efficient and flexible methods for the analysis of bacterial genomes and metagenomes. Several bio-informatics analysis methods were explored and led to the implementation of two software to support the research hypotheses: Ray Surveyor and kAAmer.

The first research hypothesis was to test the possibility of obtaining a comparison of genomes, from their simple DNA k -mers content, with results analogous to standard genomic comparisons such as average nucleotide identity or phylogenetic trees, but without the need for sequence alignments. Using Ray Surveyor software and several bacterial genomic and metagenomic analyses, we have demonstrated that it is possible to obtain such comparisons using k -mers from DNA sequences. In the study that presented the results of the research hypothesis, we also estimated the genotypic propensity of several bacterial species to clinically relevant phenotypes using specialized gene databases.

The second hypothesis was to test the possibility of developing a software for protein sequence identification, based on amino acid k -mers, which would be more efficient than existing software, specifically for the identification of proteins with a high degree of homology. The work led to the implementation of kAAmer, a software solution that allows the creation of protein databases where the sequence search is done by exact match of k -mers, while supporting sequence alignment. kAAmer has proven to be very efficient for protein sequence search with performances surpassing even the fastest sequence aligners in most scenarios. Other interest-

ing features are also offered by kAAmer, such as the possibility to host a database as a service on a permanent basis.

Finally, the third and last research hypothesis aimed to test the capacity the two software developed during the PhD project (Ray Surveyor and kAAmer) to produce viable results in a metagenomic analysis of the gut microbiota in relation to obesity. Taxonomic and functional profiling was performed with kAAmer as the *de novo* comparison of metagenomes with Ray Surveyor. The results obtained were significant and showed, among others, a trend towards higher relative abundance of the *Bacteroidetes* phylum and lower relative abundance of the *Firmicutes* and *Acinetobacteria* phyla in obese subjects. Several metabolic functions were also found to be significantly different in the normal and obese conditions, with a particular mention to the metabolism of short-chain fatty acids (SCFA) that are known to be associated with obesity.



Table des matières

Résumé	iii
Abstract	v
Table des matières	vii
Liste des tableaux	x
Liste des figures	xi
Liste des abréviations	xv
Remerciements	xix
Avant-propos	xx
Projets principaux	xx
Projets annexes	xxii
Financements	xxiv
Introduction	1
I.1 Mise en contexte	1
I.2 Séquençage de génomes	2
I.2.1 Première génération de séquençage	3
I.2.2 Deuxième génération de séquençage	4
I.2.3 Troisième génération de séquençage	7
I.3 Analyses génomique des bactéries	12
I.3.1 Préparation des lectures d'ADN	12
I.3.2 Identification des séquences biologique	14
I.3.3 Assemblage de génomes	28
I.3.4 Annotation de génomes	29
I.3.5 Comparaison de génomes par phylogénie et autres techniques	33
I.3.6 Taxonomie	36
I.3.7 Analyses prédictives en génomique	38
I.4 Analyses métagénomique des bactéries	41
I.4.1 Profilage taxonomique	41
I.4.2 Profilage fonctionnel	45
I.5 Microbiote intestinal humain et obésité	49
I.5.1 Microbiote intestinal humain	49

I.5.2	Obésité et microbiote	51
I.6	Hypothèses et approches méthodologiques	54
1	Comparaison phénétique de génomes procaryotes basée sur les k-mers	56
1.1	Résumé	57
1.2	Abstract	58
1.3	Introduction	59
1.4	Results and Discussion	61
1.5	Conclusion	73
1.6	Materials and Methods	75
1.7	Phenetic and Phylogenetic Analysis	77
1.8	Supplementary Material	79
1.9	Author Contributions	79
1.10	Acknowledgments	79
2	Base de données protéique flexible et efficace basés sur kAAmer	80
2.1	Résumé	81
2.2	Abstract	82
2.3	Main	83
2.4	Methods	86
2.4.1	Design of kAAmer	86
2.4.2	Database building	86
2.4.3	Database querying	86
2.4.4	Protein benchmark software	87
2.4.5	Other kAAmer use cases	87
2.5	Data availability	88
2.6	Code availability	88
2.7	Acknowledgements	88
2.8	Declaration	88
2.9	Figures and tables	89
3	Analyses multi-études du microbiome intestinal en relation avec l'obésité	91
3.1	Résumé	92
3.2	Abstract	93
3.3	Background	94
3.4	Methods	95
3.4.1	Cross-study samples selection	95
3.4.2	Definition of overweight and obesity	95
3.4.3	Metagenomics data preparation	95
3.4.4	Statistical analyses	96
3.4.5	Machine learning in the comparison of the microbiomes	96
3.5	Results	97
3.5.1	K-mer analyses of the metagenomes	97
3.5.2	Taxonomic analyses of the gut metagenomes	98
3.5.3	Functional analyses	100
3.5.4	Obese metagenome classification	103
3.6	Discussion	104

Conclusion et perspectives	115
Annexes	118
Bibliographie	136

Liste des tableaux

I.1	Ligne du temps non exhaustive des découvertes et technologies en lien avec la génomique. L'icône  indique une découverte ou technologie de bio-informatique alors que l'icône  englobe plutôt les découvertes issues d'un laboratoire expérimental.	11
I.2	Correspondances des deux mesures de qualités (<i>phred</i> et <i>q</i>)	12
2.1	Report of the antibiotic resistance genes identification within the pan-resistant <i>Pseudomonas aeruginosa</i> E6130952 strain from kAAmer+NCBI-arg, ResFinder and CARD databases.	90
3.1	Top 20 families based on their overall abundance in our dataset. The family codes serve as a reference for the families displayed in figure 3.2.	111
3.2	Machine learning obesity classification results from metagenomics' protein annotations	112
A1	Bases de données utilisées en génomique bactérienne	133

Liste des figures

I.1	Coût du séquençage versus le nombre de nucléotides publiés sur les bases de données GenBank et WGS du NCBI. Source des données : https://www.ncbi.nlm.nih.gov/genbank/statistics/ et https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data	2
I.2	Illustration simplifiée de la technique de séquençage de Sanger [440].	4
I.3	Réactions chimiques du pyroséquençage. Figure issue de Harrington et al. [156]	5
I.4	Technique d'amplification par différentes technologies de séquençage. Figure adaptée de Goodwin et al. [144]	7
I.5	Technique de séquençage de l'appareil SOLiD de Applied Biosystems. Figure adaptée de Voelkerding et al. [403]	8
I.6	Différentes technologies d'appareils de séquençages qui produisent de longues lectures d'ADN. Aa) Technique de séquençage <i>single-molecule real-time</i> utilisée par la technologie PacBio. Ab) Technique de séquençage nanopore utilisée dans les appareils de la compagnie Oxford Nanopore. Figure adaptée de Goodwin et al. [144]	10
I.7	Démonstration de l'algorithme <i>seed-and-extend</i> basé sur un arbre de nucléotides utilisé dans fastp. Figure issue de Chen et al. [58]	14
I.8	Démonstration d'un alignement global versus un alignement local. Figure issue du livre <i>Essentials of Bioinformatics, Volume I</i> [356].	14
I.9	Exemple d'une matrice de Needleman-Wunsch pour générer un alignement global avec des scores de +2 pour un appariement, -1 pour un mésappariement et -2 pour l'ouverture d'une brèche. Les flèches indiquent les chemins possibles pour générer l'alignement, alors que les cellules en rouge indiquent le chemin choisi pour les alignements 1 et 2. Figure tirée du livre <i>Computational Biology</i> [174].	16
I.10	Exemple d'une matrice de Smith-Waterman pour générer un alignement local. Les flèches indiquent les chemins possibles pour générer l'alignement, alors que les flèches en gras indique le traçage de l'alignement local à partir du score le plus élevé. Figure issue de Feng et al. [116]	17
I.11	Exemple d'une transformée de Burrows-Wheelers et la recherche d'une séquence dans la transformée. Figure issue de Canzar et al. [48]	20
I.12	Exemple de découpage d'une séquence en k-mers de taille 7 (7-mers) et formation d'un graphe de de Bruijn depuis les k-mers.	22
I.13	Exemple de <i>minimizers</i> utilisés dans le comptage de <i>k</i> -mers. Figure tirée de Xiao et al. [423].	24
I.14	Illustration de l'algorithme de raffinement itératif pour l'alignement de séquences multiples. Figure issue de Gotoh et al. [147].	26

I.15	Modèle HMM d'un court alignement de cinq séquences avec trois positions. Les trois positions sont représentées par un état d'appariement (carrés m), un état de délétion (cercle d) ou un état d'insertion (losanges i). Chaque carré d'appariement possède une probabilité pour chacun des acides aminés possibles à cette position. Figure issue de Eddy et al. [102].	27
I.16	Pipeline d'annotation PGAP (<i>Prokaryotic Genome Annotation Pipeline</i>) de génomes bactériens du NCBI. Voir le texte pour plus de détails sur les différentes étapes du pipeline. Figure issue de Tatusova et al. [384].	31
I.17	Diagramme pour l'identification de nouvelles espèces bactériennes. Figure tirée de Chun et al. [62]	37
I.18	Flux de travaux conseillé par Knight et al. dans son article sur les meilleures pratiques pour l'analyse de métagénomés basée sur le gène de l'ARN 16S, de métagénomique par <i>shotgun</i> et de métatranscriptomique [207].	42
I.19	Illustration des composantes qui définissent un microbiome. Figure issue de Berg et collaborateurs [27]	49
I.20	Morphologie de l'intestin humain avec micrographie électronique d'une partie de l'intestin grêle avec des bactéries identifiées en vert. Figure adaptée de Bajzer et al. [22].	50
1.1	Evaluation of simulated genome populations with Ray Surveyor. Colors and symbols represent the distance metrics used to transform the Ray Surveyor's Gram matrix into a distance matrix. Each column represents a different evolutionary distance between the genomes, based on the average branch length and bacterial species definition. Ten replicates were performed for each point. First row (A) is the cophenetic correlation between the reference phylogeny and the phenetic tree. Second row (B) is the Robinson–Foulds metric between the reference phylogeny and the Ray Surveyor derived tree.	62
1.2	Comparison of phenetic trees created using Ray Surveyor to phylogenies calculated using conserved genomes or marker genes for <i>Pseudomonas aeruginosa</i> and <i>Streptococcus pneumoniae</i> . (A) Cophenetic correlation between alignment-based phylogeny and phenetic trees calculated using four different distance metrics. (B) Fowlkes–Marlows index comparing clustering done using Ray Surveyor (correlation distance metric) and phylogeny compared with classification based on multiple locus sequence typing or serotypes.	64
1.3	Comparison of phenetic trees created using Ray Surveyor to phylogeny based on 16S gene sequence for 2,429 bacterial genomes. (A) Cophenetic correlation between alignment-based phylogeny and phenetic trees calculated using four different distance metrics. (B) Fowlkes–Marlows index comparing clustering done using Ray Surveyor (correlation distance metric) and phylogeny compared with taxonomical classification at the family rank.	67

1.4	Comparison of the relationship between strains when genome sequences are filtered using one of five filtering data sets for <i>Streptococcus pneumoniae</i> , <i>Pseudomonas aeruginosa</i> and the 2,429 representative bacterial genomes. The Heatmap represents the Canberra distance between genomes collated on a subset of k-mers. The X and Y axis of the heatmap are genomes ordered based on hierarchical clustering of the complete genome. The number in top left corner of heatmaps is the cophenetic distance, expressed in percentages, between filtered data sets and whole genome phenetic tree. The darker the shade of blue, the higher the similarity between samples.	69
1.5	Cophenetic distance between phenetic trees based on whole genome and filtered data sets for 42 bacterial species from RefSeq that included at least 100 genomes. Intensity of heatmap represents the cophenetic correlation as shown in the legend. Numbers in the heatmap are percentages of genomes with zero k-mers associated with relevant filtering data set.	70
2.1	A) Design of a kAAmer database. Three key-value stores are created within a database (K-mer Store, Combination Store, Protein Store). Colours indicate the combination (hash) value that are reused in the combination store. Proteins are numbered (p01, p02, p03) and k-mers are numbered (k01,k02,...,k08). B) Protein search benchmark. Software include blastp (v2.9.0+), ghostz (v1.0.2), diamond (v0.9.25) and kAAmer (v0.4) with and without alignment.	89
3.1	Heatmap and hierarchical clustering of the metagenome based on their shared DNA content (<i>k</i> -mers of length 31). Red color indicate close similarity between metagenome as blue color indicate more dissimilarity. BMI groups are indicated by a right triangle in the upper portion of the figure and country of the individuals on the left side by upper triangles.	106
3.2	Bacteria cladogram of the gut metagenomics data found in 640 individuals down at the family taxonomic rank. Red bars indicate higher relative abundance of the taxa in obese individuals, while blue bars indicate higher representation in normal individuals. Stars indicate that the relative abundance was significantly different in both cohorts.	107
3.3	The top 4 most abundant phyla with significant changes in obese and non-obese gut microbiota.	108
3.4	The top 9 most abundant families with significant changes in obese and non-obese gut microbiota.	109
3.5	The top 20 KEGG pathways which show the most significant changes in abundances in obese and non-obese gut microbiota.	110
A1	Exemple de rapport de résultats de fastp sur une lecture (ERR321632) d'un métagénome du microbiote intestinal humain.	119
A2	Matrice de substitution PAM250.	125
A3	Matrice de substitution BLOSUM62.	125
A4	A) Qualité de l'alignement mesuré par le score par colonne «Column Score (CS)». B) Temps d'exécution des aligneurs sur échelle logarithmique. Figure issue de Thompson et al. [388].	126
A5	Exemple d'un fichier GenBank pour le plasmide pKp199-1 de la souche <i>Klebsiella pneumoniae subsp. pneumoniae</i> CCRI-22199 [88].	127

A6	Example d'un fichier EMBL pour le plasmide pKp199-1 de la souche <i>Klebsiella pneumoniae subsp. pneumoniae</i> CCRI-22199 [88].	128
A7	Exemple d'une fonction logistique aussi nommé fonction sigmoïde.	129
A8	Example d'un fichier GFF pour le plasmide pKp199-1 de la souche <i>Klebsiella pneumoniae subsp. pneumoniae</i> CCRI-22199 [88].	130
A9	The top 6 most abundant genera with significant changes in obese and non-obese gut microbiota.	131

Liste des abréviations

Biologique :

ADN : Acide désoxyribonucléique

ARN : Acide ribonucléique

APS : Adénosine 5' phosphosulfate

ATP : Adénosine triphosphate

BLOSUM : *Block Substitution Matrix*

BWT : *Burrows-Wheeler Transform*

DDBJ : *DNA Data Bank of Japan*

ddNTP : Didésoxyribonucléotide triphosphate

ddATP : Didésoxyadénosine triphosphate

ddCTP : Didésoxycytidine triphosphate

ddGTP : Didésoxyguanosine triphosphate

ddTTP : Didésoxythymidine triphosphate

dNTP : Désoxyribonucléoside triphosphate

dATP : Désoxyadénosine triphosphate

dCTP : Désoxycytidine triphosphate

dGTP : Désoxyguanosine triphosphate

dTTP : Désoxythymidine triphosphate

EBI : *European Bioinformatics Institute*

EMBL : *European Molecular Biology Laboratory*

GFF : *General Feature Format*

Indel : Insertion ou délétion

MSA : *Multiple Sequence Alignment*

NCBI : *National Center for Biotechnology Information*

NGS : *Next Generation Sequencing*

PAM : *Point Accepted Mutation*
PCR : *Polymerase Chain Reaction*
PP_i : Pyrophosphate inorganique
SMS : *Single-Molecule Sequencing*
ZMW : *Zero-mode Waveguides*

Informatique :

API : *Application Programming Interface*
BD : Base de données
HMM : *Hidden Markov Model*
ML : *Maximum Likelihood*
NJ : *Neighbor-Joining*
NP : *Nondeterministic Polynomial time*
PCA : *Principal Component Analysis*
PCoA : *Principal Coordinate Analysis*
RAM : *Random-Access Memory*
UPGMA : *Unweighted Pair Group Method with Arithmetic mean*
SSD : *Solid-State Drive*

*À toi, dans l'ombre de mes
pensées..*

Science knows no country,
because knowledge belongs to
humanity, and is the torch which
illuminates the world.

Louis Pasteur

Remerciements

J'aimerais remercier mon directeur de thèse, le professeur Jacques Corbeil, pour m'avoir fait confiance tout au long de mon doctorat et avoir guidé mes travaux de recherches. Je remercie aussi mon codirecteur, le professeur François Laviolette, pour le partage de ses connaissances sur les algorithmes informatiques. De plus, je remercie le professeur émérite Paul H. Roy pour tous ses apprentissages sur la génétique des bactéries et pour nos nombreuses collaborations sur l'étude de la résistance aux antibiotiques.

J'aimerais aussi remercier tous les étudiants et professionnels de recherches qui ont agrémenté mes études de troisième cycle en plus d'alimenter des conversations éclairées sur les enjeux scientifiques d'aujourd'hui et de demain. Une mention spéciale au Dr Sébastien Boisvert pour son enthousiasme, sa rigueur et son support dans l'implémentation des logiciels présentés dans cette thèse. Une deuxième mention spéciale au professeur Frédéric Raymond avec qui j'ai beaucoup appris sur la métagénomique des bactéries.

Enfin, merci à toute ma famille sans qui rien de tout cela n'aurait été possible.

Avant-propos

Projets principaux

Cette thèse de doctorat est une thèse par articles où chaque chapitre constitue un article rédigé par l’auteur de la thèse, Maxime Déraspe. L’objectif principal du projet de doctorat était de contribuer aux méthodes d’analyses en génomique et en métagénomique comparatives des bactéries et de démontrer leur application via une étude sur la condition de l’obésité chez l’humain. Trois chapitres-articles sont donc présentés dans cette thèse :

- *Phenetic Comparison of Prokaryotic Genomes Using k-mers* (chapitre 1)
- *Fast protein databases as a service using kAAmer* (chapitre 2)
- *Cross-study analyses of gut microbiomes from healthy and obese individuals* (chapitre 3)

Contribution à l’article “*Phenetic Comparison of Prokaryotic Genomes Using k-mers*”

Les auteurs de l’article sont : Maxime Déraspe (MD), Frédéric Raymond (FR), Sébastien Boisvert (SB), Alexander Culley (AC), Paul H. Roy (PHR), François Laviolette (FL), Jacques Corbeil (JC).

L’implémentation du logiciel Ray Surveyor a été réalisée par MD et SB. Les algorithmes ont été conçus par MD, SB et FL. Les analyses bio-informatiques ont été réalisées par MD et FR. MD, FR, AC, PHR et JC ont interprété les résultats des analyses. MD, FR, AC et JC ont contribué à la préparation du manuscrit. Tous les auteurs ont passé en revue le manuscrit et accepté son contenu final.

Contribution à l’article “*Fast protein databases as a service using kAAmer*”

Les auteurs de l’article sont : Maxime Déraspe (MD), Sébastien Boisvert (SB), Paul H. Roy (PHR), François Laviolette (FL), Jacques Corbeil (JC).

Les algorithmes utilisés dans kAAmer ont été conçus par MD, SB et FL. L’implémentation du logiciel kAAmer a été réalisée par MD. Les analyses bio-informatiques ont été conduites

par MD et interprétées par MD, PHR et JC. MD a écrit le manuscrit et tous les auteurs ont apporté leurs commentaires et corrections.

Contribution à l'article “*Cross-study analyses of gut microbiomes from healthy and obese individuals*”

Les auteurs de l'article sont : Maxime Déraspe (MD), Charles Burdet (CB), Juan Manuel Dominguez (JMD), François Laviolette (FL), Paul H Roy (PHR), and Jacques Corbeil (JC).

La sélection des métagénomés à partir de différentes études a été concrétisée par MD et CB. Les analyses bio-informatiques sur les métagénomés ont été réalisés par MD et JMD. MD et JC ont interprété les résultats des analyses. MD a écrit le manuscrit et tous les auteurs ont apporté leurs commentaires et corrections.

Projets annexes

Durant ce doctorat, l'auteur a aussi pris part à plusieurs projets de recherches dignes de mention. Voici la liste de publications où l'auteur apparaît comme auteur ou coauteur :

- Kothari, C., Osseni, M. A., Agbo, L., Ouellette, G., Déraspe, M., Laviolette, F., Corbeil, J., Lambert, J.-P., Diorio, C., & Durocher, F. (2020). Machine learning analysis identifies genes differentiating triple negative breast cancers. *Scientific Reports*, 10(1), 10464. [212];
- Déraspe, M., Longtin, J., & Roy, P. H. (2020). Genome Sequence of a *Klebsiella pneumoniae* NDM-1 Producer Isolated in Quebec City. *Microbiology Resource Announcements*, 9(3). [88];
- Raymond, F., Boissinot, M., Ouameur, A. A., Déraspe, M., Plante, P.-L., Kpanou, S. R., Bérubé, È., Huletsky, A., Roy, P. H., Ouellette, M., Bergeron, M. G., & Corbeil, J. (2019). Culture-enriched human gut microbiomes reveal core and accessory resistance genes. *Microbiome*, 7(1), 56. [320];
- Xiong, J., Déraspe, M., Iqbal, N., Krajdén, S., Chapman, W., Dewar, K., & Roy, P. H. (2017). Complete Genome of a Panresistant *Pseudomonas aeruginosa* Strain, Isolated from a Patient with Respiratory Failure in a Canadian Community Hospital. *Genome Announcements*, 5(22). [425];
- Drouin, A., Giguère, S., Déraspe, M., Marchand, M., Tyers, M., Loo, V. G., Bourgault, A.-M., Laviolette, F., & Corbeil, J. (2016). Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics*, 17(1), 754. [98];
- Xiong, J., Déraspe, M., Iqbal, N., Ma, J., Jamieson, F. B., Wasserscheid, J., Dewar, K., Hawkey, P. M., & Roy, P. H. (2016). Genome and plasmid analysis of blaIMP-4-carrying *Citrobacter freundii* B38. *Antimicrobial Agents and Chemotherapy*, 60(11), 6719–6725. [427];
- Raymond, F., Déraspe, M., Boissinot, M., Bergeron, M. G., & Corbeil, J. (2016). Partial recovery of microbiomes after antibiotic treatment. *Gut Microbes*, 7(5), 428–434. [321];
- Raymond, F., Ouameur, A. a., Déraspe, M., Iqbal, N., Gingras, H., Dridi, B., Leprohon, P., Plante, P.-L., Giroux, R., Bérubé, È., Frenette, J., Boudreau, D. K., Simard, J.-L., Chabot, I., Domingo, M.-C., Trottier, S., Boissinot, M., Huletsky, A., Roy, P. H., ... Corbeil, J. (2015). The initial state of the human gut microbiome determines its reshaping by antibiotics. *The ISME Journal*, 1–14. [322];
- Déraspe, M., Binkley, G., Butano, D., Chadwick, M., Cherry, J. M., Clark-Casey, J., Contrino, S., Corbeil, J., Heimbach, J., Karra, K., & others. (2016). Making Linked Data SPARQL with the InterMine Biological Data Warehouse. *SWAT4LS*. [85];

- Déraspe, M., Karra, K., Binkley, G., Sullivan, J., Mickle, G., Corbeil, J., Cherry, J. M., & Dumontier, M. (2016). Semantic Research Platform for Model Organism Data. *Biomedical Data Integration and Discovery*. [87];
- Kos, V. N., Déraspe, M., McLaughlin, R. E., Whiteaker, J. D., Roy, P. H., Alm, R. A., Corbeil, J., & Gardner, H. (2015). The resistome of *Pseudomonas aeruginosa* in relationship to phenotypic susceptibility. *Antimicrobial Agents and Chemotherapy*, 59(1), 427–436. [211].

Financements

L'auteur de cette thèse de doctorat, Maxime Déraspe, fut supporté financièrement par les Fonds de recherche du Québec en Santé (#32279) et le Consortium québécois sur la découverte du médicament (CQDM). Les travaux furent aussi supportés par la chaire de recherche du Canada en génomique médicale attribuée au directeur de recherche de cette thèse, le professeur Jacques Corbeil. L'hébergement des logiciels et bases de données développés durant ce doctorat ainsi que les calculs nécessaires aux différentes analyses bio-informatique furent réalisés sur les infrastructures de Calcul Canada. Les opérations de Calcul Canada sont financées par la Fondation canadienne pour l'innovation (FCI) et supportées par les partenaires provinciaux et établissements universitaires, tels que Calcul Québec, SciNet, ACENET, Compute Ontario et WestGrid.

Introduction

I.1 Mise en contexte

Depuis l'aube de la biologie cellulaire, les chercheurs ont souhaité connaître l'origine chimique et physique des éléments moléculaires du vivant. À l'époque (~ 1930), deux champs d'intérêt étaient prêtés à la biologie cellulaire, soit l'élucidation de la génétique des microorganismes au niveau moléculaire et la structure tridimensionnelle des macromolécules biologiques, principalement les protéines [164]. Malgré le large éventail de découvertes qui s'en suivirent, mention particulière au dogme central de Francis Crick qui expliquera la relation entre l'ADN (information génétique) et sa traduction en protéine (1957) [66], un rapprochement intéressant peut être fait avec le domaine de la bio-informatique où les principaux champs d'expertise sont encore aujourd'hui la génomique, où l'on étudie le contenu cellulaire en ADN, et la structure tridimensionnelle des macromolécules biologiques (bio-informatique structurale).

La génomique a réellement été propulsée avec les avancées et la réduction des coûts dans le domaine du séquençage de l'ADN. En effet, obtenir l'ordre dans lequel apparaissent les nucléotides depuis l'ADN d'un organisme (chromosome, plasmide, etc.), constitue la base de la génomique en plus d'être un outil de prédilection pour les études génétiques liées à l'hérédité des traits. La première section (I.2) traitera plus en profondeur du séquençage de génome et des différentes technologies mises de l'avant au cours des dernières années. La section suivante (I.3) introduira les différentes méthodes d'analyse en génomique microbienne. L'accent sera porté sur les bactéries pathogènes et les analyses bio-informatiques nécessaires à l'élucidation phénotypique propre aux génomes bactériens. Par la suite, les sections I.4 et I.5 aborderont les analyses métagénomiques et la relation entre le microbiote intestinal humain et la santé humaine.

I.2 Séquençage de génomes

Le séquençage de génome a marqué une étape importante en biologie, à un point tel qu'une discipline en est née : la génomique. Plusieurs générations de séquenceurs ont permis de faire évoluer les études génomiques qui sont devenues une pratique courante dans les laboratoires de recherche et de plus en plus présente dans les soins cliniques [312]. La figure I.1 illustre la baisse du coût du séquençage en relation avec le nombre de séquences déposées dans les bases de données (BD) du NCBI¹. On peut voir une réduction importante du coût du séquençage durant la fin de la première décennie des années 2000. Les améliorations techniques du séquençage, propulsé entre autres par le projet du génome humain, et la compétition commerciale ont très certainement contribué à l'adoption générale de la génomique. L'ampleur du projet de séquençage du génome humain allait d'ailleurs ouvrir les portes à d'autres projets d'envergures, tels que ENCODE [107], HapMap [177], le projet 1000 génomes [1] et le projet du microbiome humain [278].

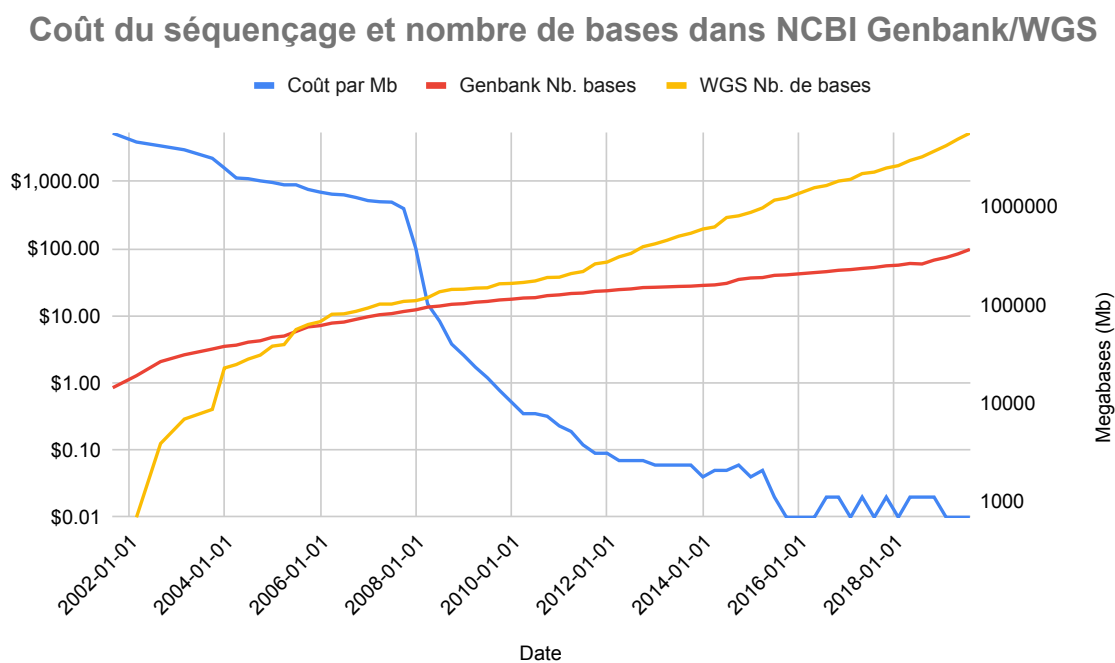


FIGURE I.1 – Coût du séquençage versus le nombre de nucléotides publiés sur les bases de données GenBank et WGS du NCBI. Source des données : <https://www.ncbi.nlm.nih.gov/genbank/statistics/> et <https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data>.

1. NCBI : National Center for Biotechnology Information

I.2.1 Première génération de séquençage

Au début des années 2000, la plupart des génomes bactériens étaient séquencés avec la méthode Sanger. La méthode fut développée par Frédéric Sanger en 1977 avec la technique de terminaison de chaîne ou dideoxy [339]. En effet, la propriété chimique de complémentarité entre les bases de l'ADN allait permettre de mesurer les bases ajoutées lors de la copie d'un simple brin d'ADN en un double brin, communément vue durant le processus de réplication d'ADN. Les déoxyribonucléotides (dNTPs : dATP, dCTP, dGTP et dTTP) sont les composantes monomères de l'ADN et ainsi l'objet à mesurer pour connaître l'encodage de la nature même du matériel génétique, soit l'ADN. Ce sont les didésoxyribonucléotides (ddNTPs : ddATP, ddCTP, ddGTP, ddTTP) qui permettent de reconstruire les séquences nucléotidiques. Les ddNTPs sont des déoxyribonucléotides dépourvus de leur partie 3' hydroxyl et qui induisent la terminaison de la synthèse du brin d'ADN. C'est à l'aide de leur radiomarquage qu'un signal peut être émis et capté pour l'identification des nucléotides. La figure I.2 illustre les éléments essentiels de la technologie de séquençage Sanger. Tout d'abord, dans quatre solutions sont mélangé une amorce (oligonucléotide de petite taille) complémentaire au brin d'ADN à séquencer, un ADN polymérase pour la synthèse de l'ADN durant la réplication du brin complémentaire, les didésoxyribonucléotides ainsi que des déoxyribonucléotides standards. Lorsqu'un didésoxyribonucléotide est ajouté au fragment d'ADN par l'ADN polymérase, l'élongation se termine à cette position alors que si un déoxyribonucléotide est ajouté alors l'élongation se poursuit. On obtient alors plusieurs fragments d'ADN de différentes longueurs qui se terminent par le ddNTP marqué et qui seront séparés par électrophorèse capillaire.

Les fluorophores des ddNTPs seront ensuite détectés avec un laser et le chromatogramme résultant sera analysé avec un logiciel. Un peu plus tard dans les années 80 (voir I.1) viendra la PCR²[338; 337] qui, combiné avec les techniques d'ADN recombinant [179; 68], permettra la production d'ADN purifié et de hautes concentrations nécessaires au séquençage. Cette première génération de séquençage produira des fragments avec des longueurs pouvant atteindre pas loin de un kilobase. Ce sont loin d'être tous les gènes d'un génome bactérien qui sont de longueur égale ou inférieure à un kilobase. L'effort nécessaire pour séquencer un génome complet avec cette technique demeure considérablement plus exigeante que les techniques actuelles. C'est en 1995 qu'était séquencé pour la première fois un génome bactérien en entier, *Haemophilus influenzae* avec une longueur de 1,830,140 paires de bases [120]. Malgré le coût élevé de la technique de Sanger, une première version du génome humain fut publiée en 2001 [73; 400].

2. Polymerase Chain Reaction

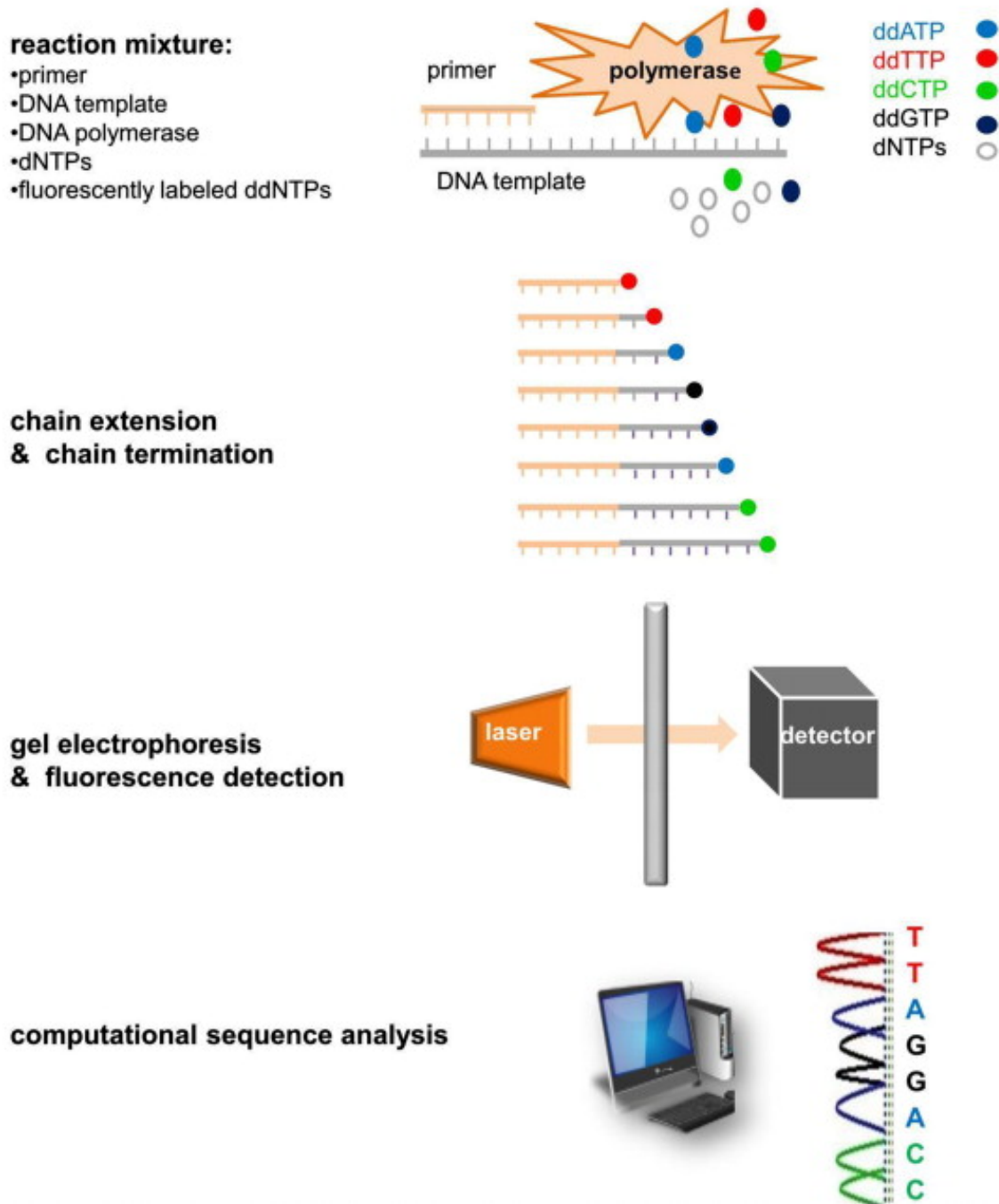


FIGURE I.2 – Illustration simplifiée de la technique de séquençage de Sanger [440].

I.2.2 Deuxième génération de séquençage

La deuxième génération (ou prochaine génération - NGS³) de séquenceur ne procédera plus par la terminaison de chaîne, mais plutôt directement en mesurant en temps réel l'activité enzymatique de l'ADN polymérase lors de la synthèse d'un brin complémentaire.

3. Next Generation Sequencing

La première technologie de la deuxième génération est le pyroséquençage qui fut introduit par Nyren et al. en 1993 [283] et qui fut amélioré en 1998 par Ronaghi et al. [333]. En 2005, une troisième évolution de la méthode par Margulies et ses collègues, associés à la compagnie *454 Life Sciences* [254], allait marquer une étape importante pour la parallélisation du séquençage et le développement d'appareils commerciaux (454 GS 20, 454 GS FLX). La figure I.3 illustre les réactions chimiques qui se déroulent durant le pyroséquençage [156]. La solution pour

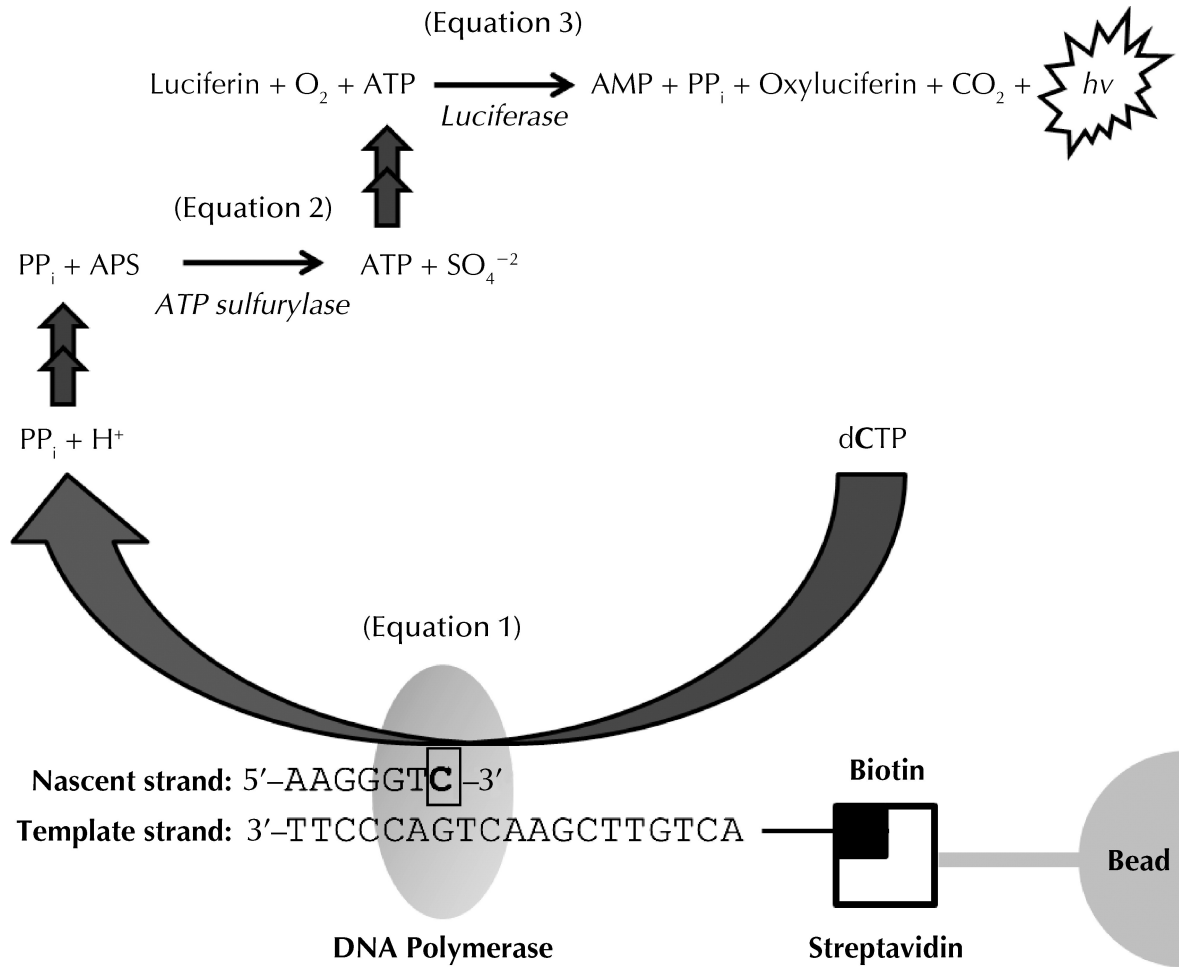


FIGURE I.3 – Réactions chimiques du pyroséquençage. Figure issue de Harrington et al. [156]

le pyroséquençage comprend l'ADN simple brin hybridé à une amorce en plus des enzymes et substrats suivants : ADN polymérase, ATP sulfurylase, luciférase, apyrase, adénosine 5' phosphosulfate (APS) et luciférine. Dans une première étape, l'ajout des dNTPs permet à l'ADN polymérase d'ajouter le bon dNTP sur le brin complémentaire d'ADN. De cette réaction est libéré un pyrophosphate (PP_i) qui est converti en ATP par l'ATP sulfurylase à l'aide du substrat APS. Ensuite, l'ATP agit comme substrat à la luciférase qui engendre la conversion de la luciférine émettant ainsi un faisceau lumineux qui sera capté par une caméra. Les dNTPs non incorporés seront dégradés par l'enzyme apyrase. Les signaux lumineux sont captés par une caméra et analysés par un programme informatique pour en déduire la séquence de nucléotides

incorporés et complémentaires au brin d'ADN séquencé. Une expérience de séquençage sur un 454 GS FLX+ durera environ une journée pour produire 1 million de lectures d'une longueur moyenne de 700 paires de bases [144].

Suite au succès du 454 et de la parallélisation du séquençage, plusieurs autres techniques firent leur apparition. L'une d'entre elles fut la technologie de Solexa qui sera par la suite acquise par la compagnie Illumina (voir Tableau I.1). En effet, dans la technique du pyroséquençage, chaque fragment d'ADN était capturé par une bille et amplifié par PCR en émulsion (I.4a). Solexa introduisit en 2006 avec leur séquenceur *Genome Analyzer* le principe de cuve de circulation (*flowcell*) (I.4b) qui sera aussi utilisé dans les appareils d'Illumina qui suivront, tels le MiSeq et le HiSeq. Avec cette méthode, des adaptateurs spécifiques sont ajoutés à l'ADN qui sera dénaturé pour leur passage sur la cuve de circulation. Les oligonucléotides complémentaires aux adaptateurs sont attachés à la surface de la cuve de circulation et capturent aléatoirement les brins d'ADN. Il y a ensuite amplification des brins d'ADN capturés par *bridge PCR* ce qui crée des agglomérats d'amplicons (100 à 200 millions de clones) qui seront aussi dénaturés pour être séquencés [144]. L'ajout de dNTPs marqués et munis d'un bloqueur réversible, des amorces et de l'ADN polymérase combiné à l'excitation par laser entraîne l'émission d'une fluorescence à l'ajout d'une base et qui est captée par une caméra. À la fin d'un cycle, les bloqueurs sur les dNTPs sont supprimés ce qui permet au prochain cycle de débiter. À titre d'exemple, la version 3 du Illumina MiSeq peut produire jusqu'à 50 millions de lectures appariées (2x) d'une longueur de 300 paires de bases, pour un rendement théorique de 15 milliards de paires de bases pour une seule expérience ayant une durée de 21 à 56 heures [144].

Une autre technologie de séquençage considéré de la deuxième génération est le SOLiD de la compagnie Applied Biosystems (I.5). Cette méthode ne procède pas par synthèse comme les appareils 454 et Illumina, mais plutôt avec un séquençage par la ligation et la détection des oligonucléotides. L'amplification des fragments d'ADN se fait aussi par PCR en émulsion comme dans le 454. C'est avec l'aide de sondes de deux bases (4 au total) avec un code de couleurs qu'il est possible de détecter les bases ajoutées lors de la ligation aux brins d'ADN de par la fluorescence relâchée durant la ligation et captée sur image. Après la détection des bases ajoutées, une phase de clivage des produits d'extensions est nécessaire pour lancer le prochain cycle de l'élongation de la séquence. Comme le code de couleur ne comprend que 4 combinaisons, il faut savoir l'identité de la première base pour inférer la deuxième. C'est lors du traitement informatique que seront déduites les positions des bases ajoutées. Aussi à noter que chaque base est séquencée deux fois pour améliorer la précision des lectures. Cependant, le débit du séquençage étant inférieur à celui des appareils Illumina (4 milliards de bases sur une durée d'environ 6 jours), la plateforme fut donc abandonnée par le fabricant Applied Biosystems et la communauté scientifique se tourna vers les séquenceurs MiSeq, HiSeq et NovaSeq de la compagnie Illumina [144].

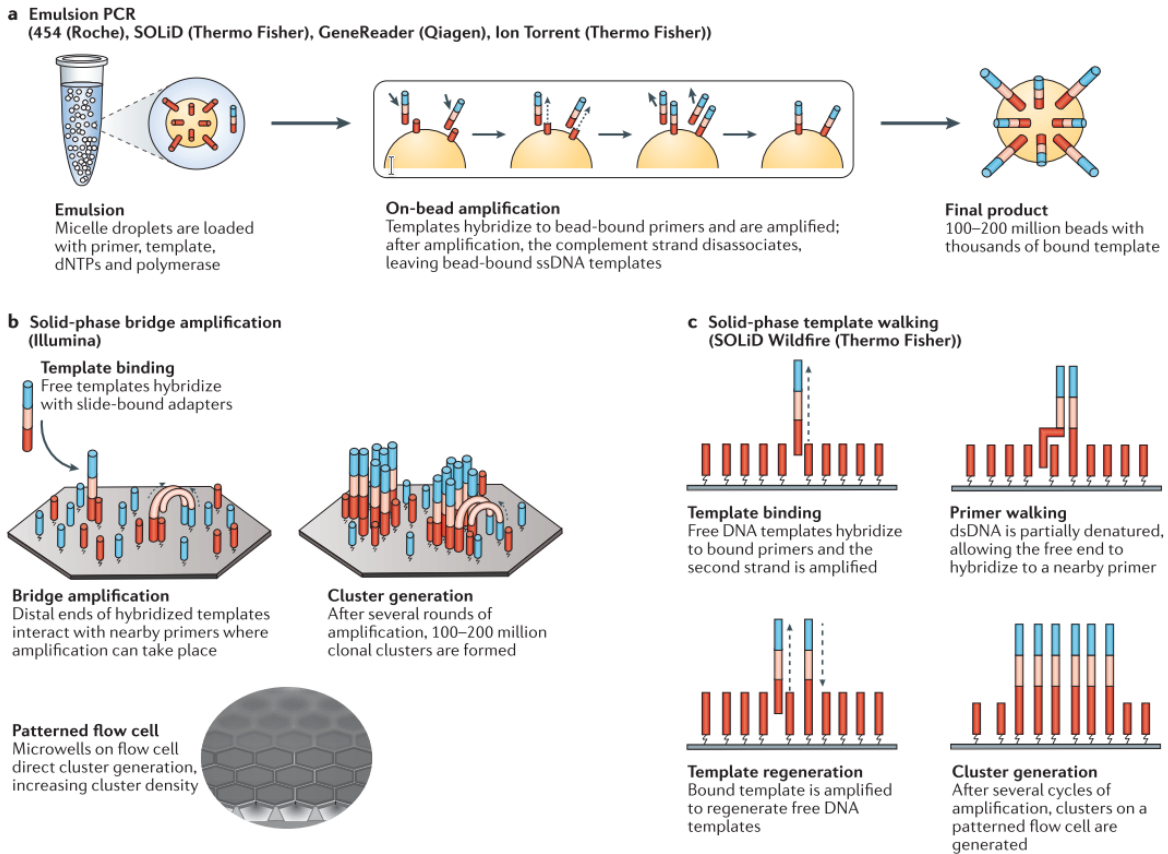


FIGURE I.4 – Technique d’amplification par différentes technologies de séquençage. Figure adaptée de Goodwin et al. [144]

I.2.3 Troisième génération de séquençage

La troisième génération de séquenceurs se définit par l’utilisation de molécules uniques d’ADN (SMS⁴) contrairement à la seconde génération où l’ADN devait être fragmenté et amplifié par PCR sur une surface solide [344]. Les deux principaux séquenceurs de la troisième génération sont le RS II de Pacific Bioscience (PacBio) et le MinION d’Oxford Nanopore. L’un des principaux avantages de la technique SMS est la longueur des lectures qu’elle génère ce qui simplifie l’assemblage de génomes *de novo* [144]. L’un des désavantages des SMS est la quantité d’erreurs que les lectures produisent [318]. Cependant, avec suffisamment de couvertures et des algorithmes adaptés pour l’assemblage, il est possible d’obtenir des séquences de très bonne qualité. Les fondations de la technique SMS proviennent du laboratoire de Stephen Quake [39; 157].

La technologie PacBio, aussi nommée séquençage en temps réel par utilisation de molécules uniques d’ADN (SMRT⁵), se base sur les travaux de Levene et al. [227] et de dispositifs à

4. SMS : *Single-Molecule Sequencing*

5. SMRT : *Single-Molecule Real-Time sequencing*

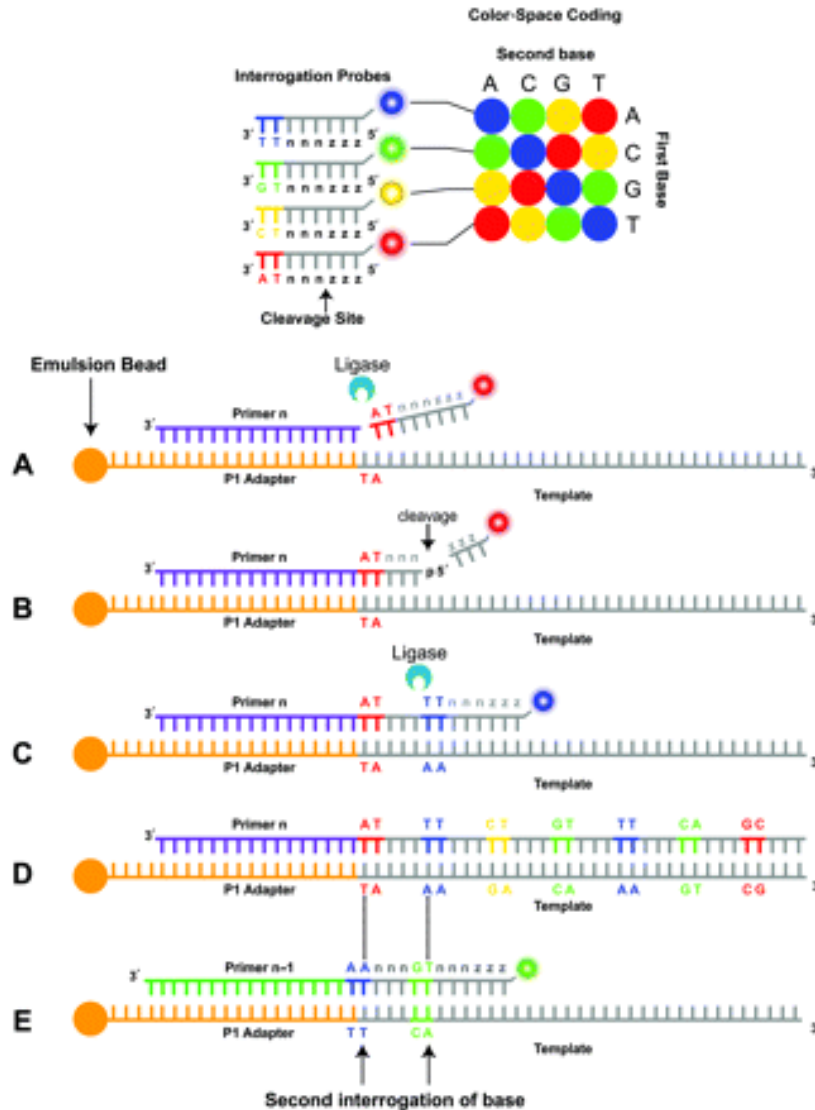


FIGURE I.5 – Technique de séquençage de l'appareil SOLiD de Applied Biosystems. Figure adaptée de Voelkerding et al. [403]

nanostructures nommés *zero-mode waveguides* (ZMW). En effet, une dizaine de milliers de ZMW sont présents sur une plaque SMRT et sur lesquels seront immobilisé un complexe muni d'une seule ADN polymérase et d'une seule molécule d'ADN simple brin. La nanostructure d'un ZMW permet donc de mesurer, à l'aide d'une source lumineuse placée sous les puits ZMW, la fluorescence produite lors de l'ajout d'un dNTP marqué et cela à l'échelle d'une seule molécule d'ADN (voir I.6 Aa). Chaque extrémité d'un fragment d'ADN possède un adaptateur avec une structure tige-boucle (ou épingle à cheveux) qui permet le séquençage circulaire en continu d'un même brin. C'est ainsi qu'un même fragment est séquençé à plusieurs reprises pour mitiger le taux d'erreurs des lectures individuelles. En effet, les lectures individuelles d'un PacBio peuvent atteindre jusqu'à 13% d'erreurs de séquençage, mais avec la technique de consensus circulaire, le taux d'erreur diminue sous la barre du 1%, soit une précision de

lecture qui se rapproche ou même surpasse la qualité des courtes lectures des technologies Illumina [380]. Le débit théorique d'un RS II de PacBio est approximativement 1 milliard de bases (55,000 lectures de 20 kilobases) sur une période de 4 heures [144].

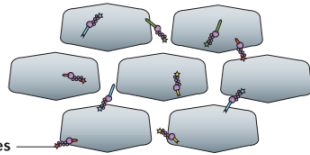
Une autre technologie populaire de la troisième génération est le MinION de la compagnie Oxford Nanopore. Ce séquenceur à l'avantage d'être très portable ne mesurant que quelques centimètres de taille et pouvant se brancher dans un ordinateur personnel standard [244]. La méthodologie utilise des nanopores, qui sont généralement des protéines transmembranaires qui permettent aux cellules de transporter des molécules aux travers leur membrane. Les protéines utilisées sont généralement issues de bactéries qui sont synthétiquement modifiées en laboratoire pour produire des pores encore plus performants pour le séquençage. Malgré le secret habituel des compagnies face à leur technologie, en 2016, Oxford Nanopore révélait qu'elle allait utiliser une version modifiée de la protéine CsgG d'*Escherichia coli* comme pore dans ses prochains appareils MinIONs. La lecture de l'ADN avec les nanopores ne requiert pas de dNTPs marqués, mais mesure plutôt la modulation des signaux électriques lors du passage de la molécule d'ADN à l'intérieur d'un nanopore. La figure I.6 Ab illustre la procédure de séquençage avec la protéine α -hémolysine qui agit comme pore. Contrairement aux fragments de la technologie PacBio, ceux d'Oxford Nanopore contiennent un adaptateur tige-boucle à une seule extrémité, pour permettre le séquençage bidirectionnel du fragment d'ADN, et l'autre extrémité porte un adaptateur qui permet de diriger le fragment vers le nanopore où il sera séquençé.

Aa Pacific Biosciences

SMRTbell template
Two hairpin adapters allow continuous circular sequencing

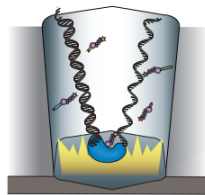


ZMW wells
Sites where sequencing takes place



Labelled nucleotides
All four dNTPs are labelled and available for incorporation

Modified polymerase
As a nucleotide is incorporated by the polymerase, a camera records the emitted light

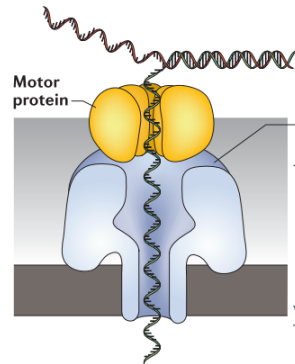
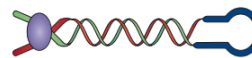


PacBio output
A camera records the changing colours from all ZMWs; each colour change corresponds to one base



Ab Oxford Nanopore Technologies

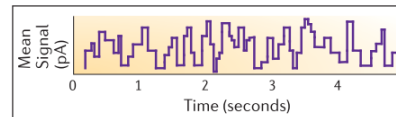
Leader-Hairpin template
The leader sequence interacts with the pore and a motor protein to direct DNA, a hairpin allows for bidirectional sequencing



Motor protein

Alpha-hemolysin
A large biological pore capable of sensing DNA

Current
Passes through the pore and is modulated as DNA passes through



ONT output (squiggles)
Each current shift as DNA translocates through the pore corresponds to a particular k-mer

FIGURE I.6 – Différentes technologies d'appareils de séquençages qui produisent de longues lectures d'ADN. Aa) Technique de séquençage *single-molecule real-time* utilisée par la technologie PacBio. Ab) Technique de séquençage nanopore utilisée dans les appareils de la compagnie Oxford Nanopore. Figure adaptée de Goodwin et al. [144]

TABLEAU I.1 – Ligne du temps non exhaustive des découvertes et technologies en lien avec la génomique. L'icône \square indique une découverte ou technologie de bio-informatique alors que l'icône \triangle englobe plutôt les découvertes issues d'un laboratoire expérimental.

1951	•	\triangle	Découverte de la séquence en acides aminées de l'insuline par Frédéric Sanger [340]
1953	•	\triangle	Découverte de la structure de l'ADN par Francis Crick et James Watson [408]
1958	•	\triangle	Découverte du dogme central par Francis Crick [74]
1970	•	\square	Conception d'un algorithme d'alignement global par Needleman et Wunsch [273]
1977	•	\triangle	Invention du séquençage par terminaison de chaîne par Frederick Sanger [339]
1981	•	\square	Conception d'un algorithme d'alignement local par Smith et Waterman [364]
1983	•	\triangle	Invention de la PCR par Kary Mullis [269]
1986	•	\triangle	Invention du premier séquenceur automatisé par la compagnie Applied Biosystems
1990	•	\triangle	Lancement du projet de séquençage du génome humain [407]
1990	•	\square	Parution de l'aligneur de séquences BLAST [8]
1993	•	\triangle	Invention du pyroséquençage par Pål Nyrén et Mostafa Ronaghi[283]
1995	•	\triangle	Séquençage complet d'un génome bactérien (<i>Haemophilus influenzae</i>) [120]
2001	•	\square	Utilisation des k -mers pour l'assemblage de génomes [305]
2001	•	\triangle	Séquençage complet du génome humain [73; 400]
2005	•	\triangle	Commercialisation du séquenceur 454 GS FLX+ par 454 Life Sciences [254]
2006	•	\triangle	Commercialisation du séquenceur Genome Analyzer par Solexa
2007	•	\triangle	Lancement du projet du microbiome humain [393]
2009	•	\square	Parution de l'aligneur de lectures BWA [232] et de l'assembleur de génome ABySS [361]
2010	•	\triangle	Commercialisation du séquenceur HiSeq 2000 par Illumina [239]
2010	•	\square	Parution de l'aligneur de lectures Bowtie [218] et des assembleurs de génomes Velvet [437] et Ray [33]
2011	•	\triangle	Commercialisation des séquenceurs MiSeq de Illumina [239] et SR de Pacific Bioscience (longues lectures) [412]
2012	•	\triangle	Commercialisation du séquenceur MinION de Oxford Nanopore (longues lectures) [54]
2012	•	\square	Parution du profileur de métagénome RayMeta [34], des assembleurs SPades [24] et MEGAHIT [230]
2015	•	\square	Parution du profileur de métagénomés MetaPhlan2 [390] et de l'aligneur de séquences DIAMOND [43]
2017	•	\square	Parution de l'assembleur de génome Canu (PacBio et Oxford Nanopore) [210]
2017	•	\triangle	Commercialisation du séquenceur NovaSeq par Illumina
2018	•	\square	Parution du profileur de métagénomés HUMAnN2 [390]

I.3 Analyses génomique des bactéries

Le rôle des séquenceurs est de produire des lectures d'ADN ou d'ARN qui seront par la suite analysées avec des outils informatiques. C'est là où la discipline de la bio-informatique entre en jeu, soit pour produire une représentation assimilable des données. Plusieurs étapes sont souvent nécessaires pour passer des lectures brutes de l'ADN, souvent de petite taille, vers une analyse complète de génomes, de population de génomes ou même de métagénomes. Les prochaines sous-sections présenteront, de manière non exhaustive, les techniques nécessaires et utilisées en bio-informatique pour mener à bien ces analyses génomiques.

I.3.1 Préparation des lectures d'ADN

La première étape pour la préparation des lectures d'ADN (*reads*) est l'identification ou l'appel des bases (*base calling*) pour la conversion des signaux électroniques du séquenceur en lectures d'ADN [61]. Le format de fichier des lectures d'ADN est le FASTQ⁶ qui comprend autant les séquences nucléotidiques que la qualité de celles-ci. Deux procédures dont il est important de tenir compte sont la suppression des adaptateurs liés aux séquenceurs pour la séparation des échantillons ainsi que la filtration des lectures par rapport à leur qualité. Cette dernière étape est étroitement liée à la technologie de séquençage utilisée. Généralement, le contrôle qualité des lectures est évalué soit par un index *phred* (q) ou une probabilité d'erreurs (p). Les deux scores sont intrinsèquement liés par la formule suivante :

$$q = -10 \cdot \log_{10}(p)$$

TABLEAU I.2 – Correspondances des deux mesures de qualités (*phred* et q)

<i>phred</i>	q
Q10	10%
Q20	1%
Q30	0.1%
Q40	0.01%
Q50	0.001%
Q50	0.0001%

Plusieurs logiciels existent pour effectuer ces contrôles qualité. Les plus populaires (pour les données Illumina) incluent entre autres cutadapt [256], Trimmomatic [35] et *fastp* [58]. Cutadapt fut initialement développé pour les données d'appareil 454, mais il fonctionne également sur des séquences Illumina. Cutadapt calcule un alignement semi-global qui ne pénalise pas pour les brèches d'alignements (*gaps*) en début et en fin de séquences. L'alignement semi-global

6. https://en.wikipedia.org/wiki/FASTQ_format

offre une flexibilité pour trouver l'emplacement idéal de l'alignement en plus de composer avec le fait que la position de l'adaptateur n'est pas connue à l'avance par l'utilisateur du logiciel. Pour le filtre de qualité, les bases à la fin des lectures qui sont en dessous du seuil de qualité spécifié seront supprimées. Trimmomatic propose deux approches pour la suppression des séquences adaptatrices aux lectures d'ADN, le mode simple et le mode palindromique. Les deux approches utilisent la technique typique d'alignement "sème et étends" (*seed-and-extend*) qui sera vue plus en détail dans la prochaine section sur l'alignement de séquences.

Au même titre que cutadapt, Trimmomatic utilise les scores de qualités des fichiers FASTQ pour couper les bases erronées des lectures d'ADN. L'algorithme utilise une fenêtre dynamique typique (*sliding window*) pour le coupage des bases erronées aux extrémités des lectures en balayant les lectures dans les sens direct (brin +, 5' vers 3') et inverse (brin -, 3' vers 5'). Lorsqu'à la moyenne au sein de la fenêtre dépasse un certain seuil d'erreur, c'est une indication à l'algorithme pour l'endroit de la coupure.

Un autre logiciel un peu plus récent (2018) pour le contrôle qualité des lectures est *fastp* qui sera utilisé dans le troisième chapitre de cette présente thèse. Le logiciel *fastp* fut développé avec comme prémisses d'être plus performant que les deux derniers logiciels présentés qui sont implémentés dans des langages de programmation considérés de plus hauts niveaux et avec un support minimal pour le *multithreading*⁷, soit Python pour cutadapt et Java pour Trimmomatic. Effectivement, *fastp* utilise le langage de programmation C++ et revendique être de 2 à 5 fois plus performant que la compétition tout en offrant plus de fonctionnalités dans un seul balayage des données de lecture. L'algorithme pour trouver les adaptateurs à une position non déterministe dans une lecture est aussi basé sur un dérivé de la technique *seed-and-extend*. L'algorithme utilise un arbre des bases nucléotidiques pour identifier l'adaptateur ainsi que les séquences en amont et en aval de celui-ci (voir figure I.7). Plusieurs analyses pertinentes sont réalisées par *fastp*, tels que l'identification des adaptateurs ou mauvaises ligations dans les lectures (A1(a)), l'identification des duplications de lectures (A1(b)), l'estimation de tailles des insertions entre les lectures appariées (A1(c)), la qualité du contenu en bases et le compte de *k*-mer avant et après filtres (A1(e,f)). Les résultats de *fastp* sont aussi présentés de manière conviviale, supportant même le HTML comme format de sortie.

Lorsque les lectures ont la qualité désirée, on procède alors à leur analyse. Il existe différentes approches pour l'analyse des séquences biologiques en génomique bactérienne. Certaines approches procèdent directement avec l'identification des gènes ou taxa associés dans les lectures d'ADN depuis les fichiers FASTQ. D'autres approches utilisent plusieurs étapes pour l'identification des gènes, qu'on identifie souvent comme des pipelines d'annotations. La prochaine section traitera plus en profondeur des différents algorithmes d'alignement souvent utilisés en bio-informatique. La section suivante traitera de l'assemblage de génome, qui peut être considéré comme une autre approche plus indirecte dans l'analyse des lectures de séquences

7. Multithread : programme en chapelet.

d'ADN.

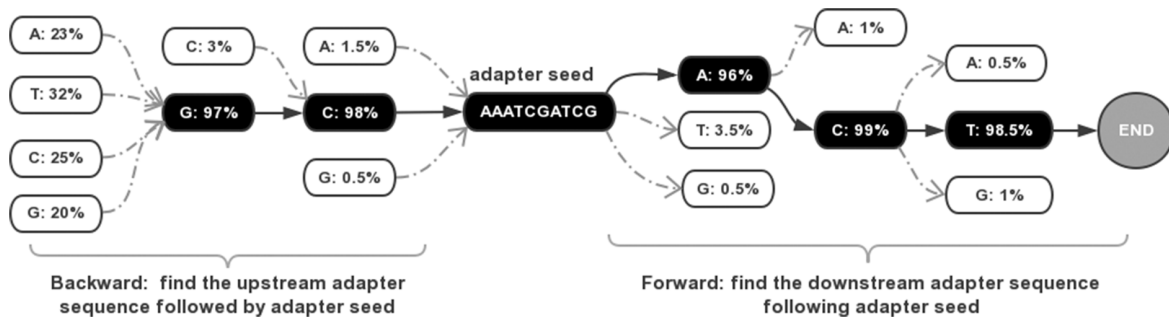


FIGURE I.7 – Démonstration de l’algorithme *seed-and-extend* basé sur un arbre de nucléotides utilisé dans fastp. Figure issue de Chen et al. [58]

I.3.2 Identification des séquences biologique

L’une des étapes primaires et primordiales à la compréhension de données génomiques (séquences d’ADN ou d’ARN) est l’identification des séquences nouvellement obtenues depuis leur comparaison avec des séquences déjà connues et caractérisées. La principale méthode pour y arriver est l’alignement de séquences, mais il existe aussi des techniques dites sans alignement (*alignment-free*) qui ont été développées pour contrer le temps d’exécution non négligeable d’un alignement de séquences.

Identification de séquences avec alignement

L’alignement de séquences peut être divisé en trois catégories, soit l’alignement par paire, l’alignement multiple (MSA⁸) et l’alignement à un profil de séquence [60; 61].

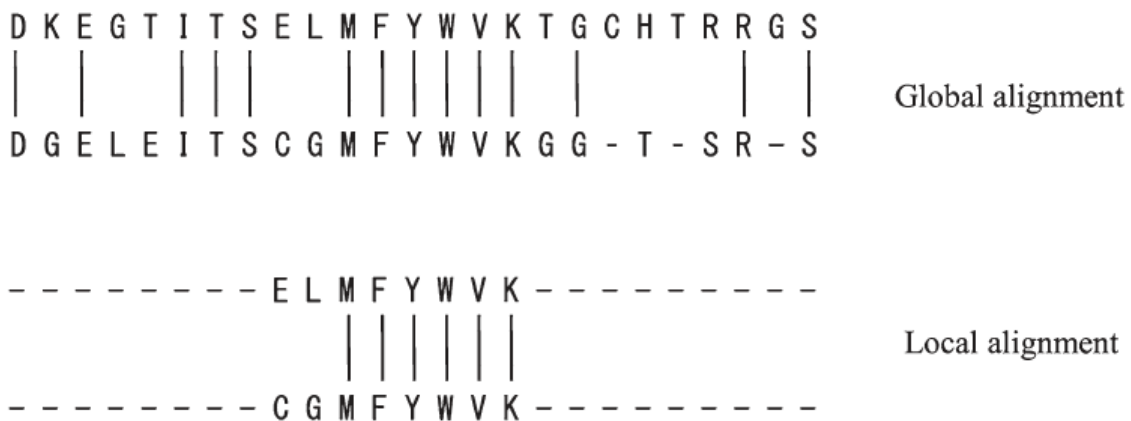


FIGURE I.8 – Démonstration d’un alignement global versus un alignement local. Figure issue du livre *Essentials of Bioinformatics, Volume I* [356].

8. MSA : *Multiple Sequence Alignment*

D'abord, l'alignement par paire ou deux à deux consiste à aligner deux séquences l'une contre l'autre. Deux méthodes sont envisageables lors d'un alignement par paire, soit celle de l'alignement global ou de l'alignement local. La figure I.8 démontre la différence que peuvent apporter les deux approches d'alignements.

L'alignement global consiste à découvrir le meilleur alignement possible pour chacun des nucléotides (ADN) ou acides aminés (protéine) sur la pleine longueur des deux séquences à aligner. Le premier algorithme de programmation dynamique destiné à l'alignement global de séquences fut développé en 1970 par Needleman et Wunsch [273]. La programmation dynamique est essentiellement l'art de diviser un problème complexe en sous-problèmes plus simple. Les principales étapes pour résoudre l'alignement global avec l'algorithme de Needleman-Wunsch sont :

- la création d'une grille avec comme index la première séquence et comme entête la deuxième séquence ;
- le choix d'une matrice de scores pour l'appariement des bases ou des acides aminés et les pénalités pour l'ouverture et l'extension des brèches dans l'alignement ;
- le remplissage de la grille avec les scores correspondant aux matrices de substitution et aux pénalités ;
- la résolution des chemins dans la grille avec les meilleurs pointages en partant de la fin des deux séquences, soit le coin inférieur droit de la grille ;
- la présentation des alignements selon les meilleurs chemins (pointage) trouvés dans la grille.

La figure I.9 présente le grille résultante d'un alignement global qui utilise la technique de Needleman-Wunsch.

L'alignement local était à l'origine surtout destiné à aligner les séquences de différentes longueurs. Le premier algorithme de programmation dynamique pour l'alignement local fut développé par Smith et Waterman en 1981 (Smith-Waterman) [364]. À l'instar de l'algorithme de Needleman-Wunsch, une grille avec les deux séquences est construite pour la résolution de l'alignement. Cependant, pour l'initialisation des scores, la première ligne et la première colonne ne sont pas sujettes aux brèches et sont plutôt fixées à zéro. Similairement, les scores ne peuvent être négatifs comme dans Needleman-Wunsch et seront aussi fixés à zéro. Pour le traçage du chemin au lieu de prendre exactement le coin inférieur droit de la grille, c'est plutôt le plus haut pointage qui sera choisi et étendu jusqu'à la rencontre d'une cellule avec un score de zéro (voir figure I.10). La complexité algorithmique pour Smith-Waterman et Needleman-Wunsch est de $O(mn)$ (m et n correspondent aux longueurs des séquences) dans le pire scénario, autant en termes de performance que d'espace requis.

L'un des algorithmes les plus populaires pour l'alignement local est assurément NCBI BLAST (*Basic Local Alignment Search Tool*) [8]. BLAST utilise une méthode de recherche pour ap-

		M	V	S	S	D
	0	-2	-4	-6	-8	-10
M	-2	2	0	-2	-4	-6
V	-4	0	4	2	0	-2
S	-6	-2	2	6	4	2
D	-8	-4	0	4	5	6

Alignment 1:

M	V	S	S	D
M	V	S	-	D

Alignment 2:

M	V	S	S	D
M	V	-	S	D

FIGURE I.9 – Exemple d’une matrice de Needleman-Wunsch pour générer un alignement global avec des scores de +2 pour un appariement, -1 pour un mésappariement et -2 pour l’ouverture d’une brèche. Les flèches indiquent les chemins possibles pour générer l’alignement, alors que les cellules en rouge indiquent le chemin choisi pour les alignements 1 et 2. Figure tirée du livre *Computational Biology* [174].

proximer le score qui serait reporté par un alignement exhaustif comme Smith-Waterman tout en améliorant l’efficacité de la recherche [61]. En effet, la performance devient un enjeu lorsqu’on veut permettre la recherche dans des bases de données qui contiennent un nombre important de séquences. L’heuristique de BLAST exploite des mots ou k -tuples de tailles fixes (habituellement 3 acides aminés ou 11 nucléotides) pour rapidement filtrer les séquences candidates dans la base de données. Plusieurs logiciels d’alignements (aligneurs) offrent des coffres d’outils similaires à BLAST pour différents types d’alignement avec différentes variantes algorithmiques. Pour en nommer quelques-uns, il existe BLAT [202], MUMmer [214], USEARCH [104], DIAMOND [43] et FASTA [300]. Tout comme dans Smith-Waterman, BLAST utilise des matrices de substitution pour le calcul des scores d’alignements. Pour l’ADN, les valeurs par défaut actuellement utilisé par BLAST du NCBI sont de 1 point pour un appariement et de -2 points pour un mésappariement de bases. Les matrices de substitutions d’acides aminés les plus connues pour l’alignement de protéines sont certainement PAM250 (*Point Accepted*

		Database sequence							
		T	C	T	C	G	A	T	
Query sequence	G	0	0	0	0	0	0	0	0
	T	0	→ 0	→ 0	→ 0	→ 0	→ 8	→ 7	→ 6
	C	0	→ 4	→ 18	→ 17	→ 18	→ 17	→ 16	→ 15
	T	0	→ 5	→ 17	→ 23	→ 22	→ 21	→ 20	→ 21
	A	0	→ 4	→ 16	→ 22	→ 20	→ 22	→ 26	→ 25
	C	0	→ 3	→ 17	→ 21	→ 35	→ 34	→ 33	→ 32

FIGURE I.10 – Exemple d’une matrice de Smith-Waterman pour générer un alignement local. Les flèches indiquent les chemins possibles pour générer l’alignement, alors que les flèches en gras indique le traçage de l’alignement local à partir du score le plus élevé. Figure issue de Feng et al. [116]

Mutation 250 [79] et BLOSUM62 (*Block Substitution Matrices 62*) [162] (voir figures annexes A2 et A3). Les matrices de substitutions sont symétriques, c’est-à-dire qu’elles sont carrées et que chaque position $a_{i,j} = a_{j,i}$. Il est donc possible d’utiliser seulement le triangle inférieur ou supérieur pour obtenir l’ensemble de son information. Chacune des entrées dans les matrices correspond au score, une probabilité logarithmique donnée pour un appariement des deux acides aminés (ligne / colonne) dans un alignement. La création des matrices de substitutions PAM et BLOSUM sont similaires du fait qu’ils évaluent la probabilité de substitutions entre les acides aminés basés sur l’alignement de séquences homologues connues. Par exemple, la matrice PAM250 est dérivée de 71 ensembles d’alignements globaux de protéines conservées à 85% d’identités et compte un total d’environ 1,500 mutations [357]. Le terme «point accepted mutation» signifie qu’une mutation ponctuelle d’un acide aminé ne change pas la fonction de la protéine et est ainsi tolérée par la sélection naturelle. En contrepartie, les matrices de substitutions BLOSUM sont construites à partir d’alignements multiples locaux (et non globaux comme avec PAM). Les alignements sont construits à partir d’ensembles de séquences avec un seuil d’identités données qui constitue les blocs de substitutions qui serviront à la construction de la matrice. Par exemple, si les séquences d’un bloc partagent entre elles un pourcentage d’identités de 62%, la matrice de substitution dérivée sera BLOSUM62. BLOSUM62, publié par Henikoff et Henikoff en 1992, est toujours à ce jour l’une des matrices

de substitutions les plus populaires. En effet, à l'écriture de cette thèse, elle est encore utilisée par défaut dans NCBI BLASTp avec des pénalités de 11 points pour la création d'une brèche dans l'alignement et de 1 point pour leur extension. Les résultats de scores d'alignement reportés par BLAST sont principalement le pourcentage d'identités, *bit-score* et une *E-value*. Le pourcentage d'identités est simplement le nombre d'acides aminés / nucléotides conservés sur la longueur de l'alignement. Un résultat intermédiaire de BLAST pour le calcul des scores est le score brut S , qui est calculé en fonction de la matrice de substitution et des pénalités obtenues avec les brèches d'alignements. De ce résultat sont ensuite extraits les deux autres scores mentionnés. Le *bit-score* est la version normalisée du score brut S sur un échelle logarithmique et la formule suivante est utilisée pour son calcul (K et λ sont des constantes précalculées⁹ pour chaque matrice de substitution) :

$$S' = \frac{\lambda S - \ln K}{\ln 2} \quad (1)$$

Le *E-value* donne une indication de la probabilité d'obtenir l'alignement par chance, s'il n'en tenait que du hasard. L'équation suivante permet de calculer sa valeur en tenant compte de la longueur de la séquence (m) et de l'espace de recherche dans la base de données de séquences (n) :

$$E = Kmne^{-\lambda S} \quad (2)$$

Par souci de brièveté, tous les algorithmes ne seront pas présentés en détail dans cette thèse et le lecteur est prié de se référer aux articles correspondants pour de plus amples informations. Néanmoins, attardons-nous à l'algorithme de DIAMOND [43] qui revendique être jusqu'à 20,000 fois plus performant que BLASTX tout en gardant un degré de sensibilité similaire. DIAMOND introduit la notion de double indexage pour la recherche de protéines au sein d'une base de données, autant à partir de séquences d'ADN (BLASTX) que de séquences de protéines (BLASTP) pour les requêtes. Dans les bases de données de séquences traditionnelles, comme celle construite avec BLAST, les *seed* des séquences sont indexées pour accélérer les appariements entre les séquences requêtes et les protéines dans la BD. Le double indexage de DIAMOND consiste en plus à indexer les *seed* des séquences requêtes pour accélérer la routine d'identification des séquences candidates dans la BD. De ce fait, l'algorithme se voit améliorer en performance avec la quantité de requêtes introduites. Le lecteur est prié de se référer au troisième chapitre pour l'illustration de la performance de DIAMOND avec une quantité variable et croissante de requêtes en protéines (voir figure 2.1B)). Ce même chapitre (2) introduira aussi un nouveau logiciel qui permet la création de bases de données de protéines et la recherche dans celle-ci avec ou sans alignement.

9. Voir <https://github.com/zorino/kaamer/blob/master/pkg/align/matrixScores.go>

D'autres techniques d'alignement ont aussi été développées avec l'apparition des courtes lectures des appareils de séquençage de nouvelle génération, comme celles d'Illumina. Les courtes lectures ont leur particularité et ont fait l'objet de développement logiciel important principalement pour leur mappage sur des génomes de références, souvent utilisé avec les génomes d'organismes supérieurs (humain, souris, etc.) [48]. Tout d'abord, dans les points à tenir compte s'insère les erreurs de séquençage et la possible divergence entre le génome de référence et l'organisme séquençé. Il faut donc des aligneurs capables de tolérer une quantité de mésappariements tout en déterminant l'endroit le plus probable de la lecture sur le génome de référence. Les aligneurs devraient aussi offrir la possibilité de tenir compte de l'information des lectures appariées ou des particularités des lectures selon le séquenceur utilisé. Canzar et al. introduisent deux méthodes importantes dans l'analyse et le mappage de courtes lectures sur des génomes de références, soit celles basées sur des techniques de hachage et celles basées sur la transformée de Burrows-Wheeler [48].

Les techniques de hachage consistent à utiliser une table de hachage pour stocker et indexer les k -mers, soit du génome de référence ou encore des courtes lectures. Une table de hachage est une structure de données en informatique qui implémente le type abstrait de tableau associatif de clés et de valeurs où chacune des clés est associée à une seule valeur et l'accès à celle-ci se fait en moyenne en $O(1)$. Habituellement, les tables de hachage pour le mappage de lectures auront comme clé les k -mers et comme valeur leur position dans le génome de référence ou dans les lectures analysés. La première étape algorithmique consistera donc à créer la table de hachage à partir des séquences cibles. Ensuite, la table de hachage sera interrogée pour localiser les k -mers dans le génome de référence par exemple. Finalement, un alignement généralement approximatif avec le modèle *seed-and-extend* sera réalisé sur les meilleures séquences candidates obtenues avec la table de hachage. Dans ce cas, on peut déduire que la table de hachage agit comme filtre pour rapidement éliminer les séquences non prometteuses dans l'identification des lectures. Plusieurs logiciels de mappage de courtes lectures utilisent cette technique, pour certains ce sont les lectures qui sont indexées, tels que dans Eland (Illumina non publié¹⁰), mrsFAST [155], MAQ [234], RMAP [363], ZOOM [438], SeqMap [184] et SHRiMP [336], et pour d'autres c'est le génome de référence qui est indexé, tels que dans SOAP [148], Novoalign (novocraft non publié¹¹), Mosaik [225], SRmapper [143], Stampy [247], BFAST [169], Hobbes [2] et FastHASH [424].

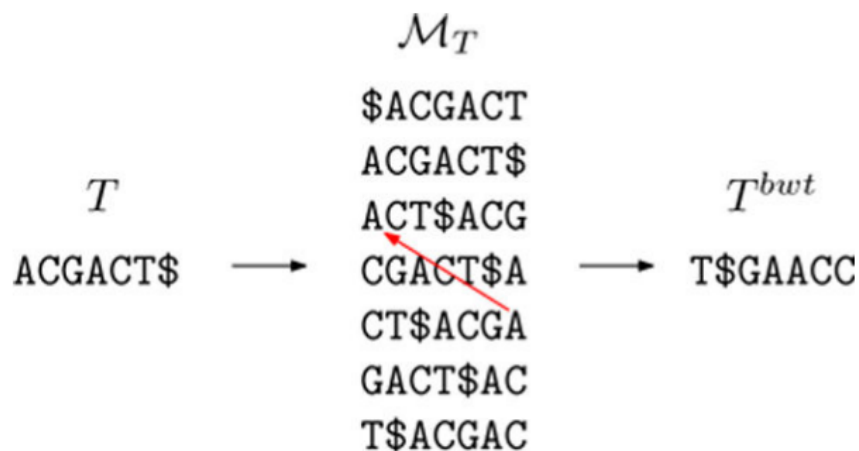
La transformée de Burrows-Wheeler (BWT) est vraisemblablement la technique la plus utilisée dans l'alignement des courtes lectures. Les logiciels Bowtie [218; 219] et BWA [232; 233] sont deux des aligneurs les plus populaires qui utilisent la technique BWT. Essentiellement, la transformée de BWT est une permutation réversible d'une chaîne de caractères qui facilite la compression des données étant donné l'ordre lexicographique dans lequel il représente la

10. https://www.illumina.com/documents/products/datasheets/datasheet_genomic_sequence.pdf

11. <http://www.novocraft.com>

chaîne de caractères. À titre d'exemple, la BWT est utilisée dans le logiciel de compression bzip2 [355]. De par la nature répétitive et massive des séquences biologiques, une méthode permettant la compression des données avec une accession à faible coût est très appropriée. La figure I.11(a) illustre la construction d'une matrice BWT et de l'encodage d'une courte séquence d'ADN avec en (b) la recherche d'une sous séquence au sein de la BWT. Au coeur de l'utilisation de BWT, dans l'alignement de séquence, se trouve l'index FM (Ferragina et Manzini [118]) qui permet l'identification rapide d'appariement de séquences exactes en agissant similairement à un tableau de suffixes. Un avantage de l'index FM est qu'il permet l'indexation et la compression des données simultanément.

(a) Transformée de Burrows-Wheelers sur la séquence ACGACT\$.



(b) Recherche de la séquence GAC au sein de la transformée de Burrows-Wheeler de la séquence ACGACT\$.

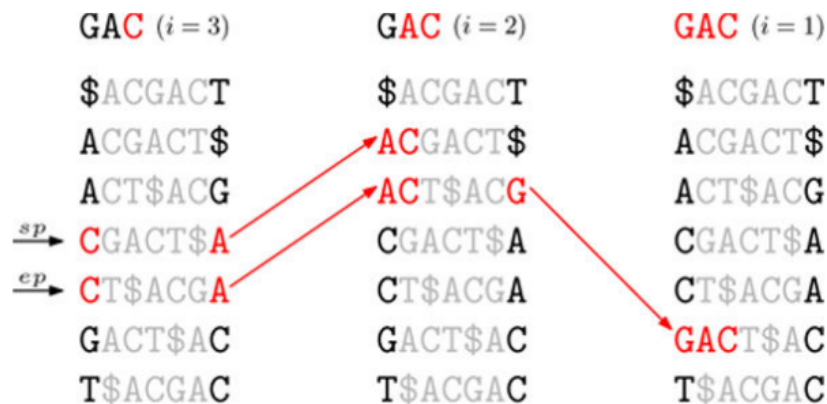


FIGURE I.11 – Exemple d'une transformée de Burrows-Wheelers et la recherche d'une séquence dans la transformée. Figure issue de Canzar et al. [48]

La prochaine sous-section présentera une alternative à l'alignement pour l'identification de

séquences ainsi que différents outils et méthodes utilisés.

Identification de séquences sans alignement

Zielezinski et al. [442] font mention de cinq cas où l'alignement peut s'avérer problématique. Premièrement, l'alignement assume une colinéarité dans les séquences alignées qui peut être enfreinte dans des scénarios réels de l'évolution des gènes. Par exemple, les génomes des virus démontrent une grande variabilité à cause de leur évolution et adaptation rapide. Ils sont fréquemment exposés à la duplication ou délétion de gènes, aux transferts horizontaux ou aux évènements de recombinaisons en plus d'avoir dans certains cas un taux de mutations élevé. Deuxièmement, la précision des alignements chute drastiquement lorsque l'homologie des séquences diminue sous un certain seuil. Notamment en pratique, les protéines qui ne partagent pas plus que 20% à 35% d'identités sont considérées dans une zone grise d'homologie. Troisièmement, les approches d'alignement sont généralement gourmandes en termes de consommation en mémoire vive (RAM¹²) et de temps de calcul requis pour compléter le *seed-and-extend*. Quatrièmement, du point de vue de la complexité algorithmique, un alignement multiple de séquences (I.3.2) est un problème NP-difficile, soit qui ne peut être résolu de façon déterministe en temps polynomial par rapport à la taille de l'entrée (nombre de séquences). Ce dernier point explique pourquoi il y a toujours de l'intérêt dans la recherche de nouvelles solutions algorithmiques et computationnelles pour ce problème qui est vieux de plus de 40 ans. Finalement, un alignement se base sur plusieurs hypothèses à propos de l'évolution des séquences comparées. L'utilisation des matrices de substitutions, les scores de pénalités préétablis et les seuils de confiance sont des variables qui ont un impact important sur le résultat des alignements. Introduisons donc quelques techniques de comparaison de séquences qui n'utilisent pas d'alignement (*alignment-free*).

L'approche des k -mers sera aussi présentée dans la section I.3.3 pour l'assemblage de génomes et elle est utilisée dans le logiciel kAAMer, présenté au chapitre 2. De façon générale avec l'utilisation de k -mers, les séquences doivent être découpées en sous-séquences qui se chevauchent avec une longueur prédéfinie, soit la longueur des k -mers (21-mers, 31-mers, etc.). Voir la figure I.12 pour une illustration du découpage en k -mers (7-mers) d'une séquence d'ADN. Ensuite, une approche typique est d'extraire les comptes de k -mers de la séquence analysée et les mettre sous forme de vecteurs. Enfin, une mesure de similarité ou de distance peut être calculée entre ces vecteurs pour la comparaison des séquences. Dans la figure 1.1 du chapitre 1 sont comparées différentes mesures de distances avec différentes longueurs de k -mer et leur corrélation avec une phylogénie de référence basée sur des données génomiques simulées. La longueur de k -mer est vraisemblablement la valeur la plus critique pour la comparaison de séquences biologiques. En effet, plus la longueur de k -mer est petite, plus la probabilité d'avoir un appariement de k -mer par chance est élevée. À l'opposé, plus la taille de k -mer

12. RAM : *Random-Access Memory*

Sequence

ATGGAAGTCGCGGAATC

7-mers

(k-mers of length 7)

ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGC
AGTCGCG
GTCGCGG
TCGCGGA
CGCGGAA
GCGGAAT
CGGAATC

de Bruijn graph

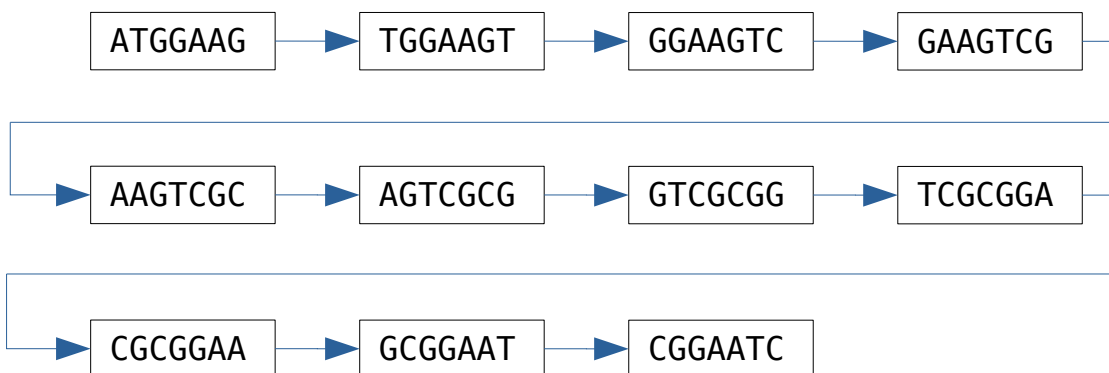


FIGURE I.12 – Exemple de découpage d’une séquence en k-mers de taille 7 (7-mers) et formation d’un graphe de de Bruijn depuis les k-mers.

est grande et moins probable l’appariement par chance devient, mais il y aura aussi moins de signaux pour les séquences avec une plus grande distance évolutive. Donc, le choix de la taille des k -mers peut être vu comme un compromis entre la spécificité et la sensibilité et aura un impact différent selon la distance évolutive entre les génomes comparés [89]. Souvent pour optimiser l’assemblage de génomes plusieurs longueur de k -mers seront utilisées, telles que dans l’assembleur SPades [24]. Il existe aussi plusieurs variantes de l’utilisation de k -mers pour la comparaison de génomes. Par exemple, une variante intéressante est utilisée dans Mash, où les k -mers sont extraits pour produire une signature (*sketch*) de génome à des fins de génomique comparative [289]. En effet, la technique MinHash¹³ est utilisée dans l’outil Mash pour la comparaison rapide de génomes, avec une preuve de concept présentée pour les génomes bactériens. La technique MinHash sera vue un peu plus en détail dans la section I.3.5.

13. MinHash : *min-wise independent permutations locality sensitive hashing*

La technique de comptage de k -mers est vraisemblablement l'une des approches les plus fondamentales de l'utilisation des k -mers en bio-informatique [83]. Cependant, plusieurs logiciels ont implémenté des techniques de programmation très avancées pour les rendre performants et efficaces. Il existe plusieurs logiciels de compteurs de k -mers, dont KMC2 [83], Jellyfish [251], MSPKmerCounter [235] et DSK [326].

Une des techniques souvent employées avec les compteurs de k -mers est la notion de *minimizer*. Les *minimizers*, tel qu'introduit par Roberts et al., ont pour fondation de diminuer l'espace requis (sur disque ou en mémoire) pour stocker les séquences biologiques [327]. En effet, au lieu de stocker tous les k -mers distincts des séquences nucléotidiques, seulement ceux qui sont représentatifs seront gardés pour être utilisés comme *seed* pour la recherche de séquences. Les k -mers gardés, nommés *minimizers*, sont généralement les plus petits lexicographiquement sur une fenêtre donnée de longueur w . La figure I.13 présente les (w,k) -minimizers des 5-mers d'une séquence de nucléotides ("GTCACGCACGTCA") avec $w=3$ et $k=3$ [423]. On constate que plusieurs k -mers consécutifs possèdent le même *minimizer* et c'est cette propriété qui permet d'économiser en mémoire et en stockage informatique par les logiciels qui utilisent cette technique. À titre d'exemple, MSPKmerCounter utilise les *minimizers* pour partitionner les données de séquences, une technique qu'ils nomment «*minimum substring partitioning*» [235], et KMC2 [83] qui utilise un sous-ensemble de *minimizers*, qu'ils nomment *signature*, pour améliorer l'efficacité de leur outil.

Alignement Multiple de Séquences (MSA)

L'alignement multiple de séquence ou MSA consiste à trouver le meilleur alignement possible pour un ensemble de séquences homologues [357]. Les MSA sont entre autres utilisés pour la construction d'arbres phylogénétiques et du fait même pour l'étude évolutive des gènes ou protéines. Le problème d'alignement multiple est généralement vu comme un problème NP-complet, c'est-à-dire qu'il n'existe pas de méthode connue pour obtenir un résultat optimal en temps polynomial. Il existe plusieurs méthodes pour réaliser des MSA avec différents niveaux de précision et de complexité. Introduisons brièvement quelques-unes de ces méthodes (alignement progressif, alignement itératif, modèle de Markov caché (HMM¹⁴) et algorithme génétique) étant donné leur rôle souvent essentiel en génomique comparative.

Une des premières méthodes progressives est celle de Feng et collaborateurs, qui part de l'hypothèse qu'un haut degré de similarité entre les gènes indique une relation évolutive entre ceux-ci [114]. La première étape de l'alignement progressif introduite par Feng et al. est de réaliser un alignement global des séquences avec l'algorithme de Needleman-Wunsch et d'en ressortir un arbre guide (*guide-tree*) ou phylogénétique pour orienter le reste de l'alignement multiple. L'arbre est généralement construit avec une méthode de *clustering* (agrégation) simple comme

14. HMM : *Hidden Markov Model*

Sequence	GTCACGCACGTCA
5-mers with each (3, 3)-minimizer in Bold	GTCAC TCACG CACGC ACGCA CGCAC GCACG CACGT ACGTC CGTCA

FIGURE I.13 – Exemple de *minimizers* utilisés dans le comptage de *k*-mers. Figure tirée de Xiao et al. [423].

du *neighbor-joining* (NJ) ou encore du UPGMA¹⁵ et non avec des méthodes de phylogénie plus sophistiquées basées sur des méthodes bayésiennes ou de maximum de vraisemblance (ML¹⁶). En se basant sur l'arbre guide, les deux branches les plus rapprochées seront choisies pour en ressortir une séquence consensus depuis leur alignement et cette dernière deviendra une représentation de la paire de branches en question. Ensuite de manière progressive, les deux prochaines branches les plus rapprochées dans l'arbre guide seront choisies pour l'extraction d'une séquence consensus, et ainsi de suite. Les alignements peuvent donc être fait entre séquences consensus ou séquence individuelle jusqu'à ce que toutes les branches soient rajoutées au consensus de l'alignement multiple. Un des avantages de l'alignement multiple progressif inclut la rapidité d'exécution en comparaison avec les méthodes de programmation dynamique. Cependant, de par son heuristique qui utilise un arbre guide, la technique est sensible à l'ordre d'inclusion des séquences dans le consensus et il y a un risque de propagation des erreurs de par la nature progressive de l'algorithme. Néanmoins, il existe plusieurs logi-

15. UPGMA : *Unweighted Pair Group Method with Arithmetic mean*

16. ML : *maximum likelihood*

ciels d’alignement multiple très populaires, tels que Clustal [220; 359], T-Coffee [280], MAFFT [198] et MUSCLE [103] qui utilisent cette technique progressive. Différentes heuristiques à la méthode originale sont introduites par ces aligneurs, soit pour améliorer la performance ou encore la précision des alignements multiples, mais ne seront pas couvertes en détail dans cette thèse. Le lecteur peut se référer à la figure annexe A4 pour une comparaison de la précision et de la performance des différents aligneurs tout juste mentionnés.

Pour les méthodes d’alignement multiples dites itératives, attardons-nous d’abord sur le raffinement itératif. La méthode de raffinement itératif introduite par Gotoh et al., revendique une amélioration significative de la qualité des MSA particulièrement pour les séquences plus éloignées évolutivement, c’est-à-dire avec un faible degré d’homologie [147]. La figure I.14 illustre la procédure utilisée dans le raffinement itératif. Similairement aux méthodes progressives, les alignements itératifs se basent sur un arbre guide pour obtenir l’ordre des séquences à aligner. Par contre, plusieurs cycles de correction de l’alignement seront réalisés pour optimiser l’alignement en maximisant une fonction objectif et cela jusqu’à convergence de l’alignement. La fonction objectif est généralement la somme des paires (*sum-of-pairs*) ou un dérivé, telle qu’une somme des paires pondérée. Cela vient contrer la dépendance de l’ordre d’ajout des séquences dans les méthodes progressives, où une séquence ajoutée à l’alignement multiple ne sera jamais revisitée. En théorie, les itérations serviront à améliorer la précision des alignements avec comme inconvénient l’augmentation de la complexité algorithmique comparativement aux alignements progressifs.

La somme des paires peut être vue comme la somme des scores d’alignement entre chacune des paires des séquences de l’alignement multiple. Les scores d’alignements (*COST*) se basent normalement sur les matrices de substitutions (PAM ou BLOSUM) combinées aux pénalités pour les brèches d’alignements, tel que discuté dans la sous-section I.3.2. Une somme des paires pondérée tiendra compte de la distance évolutive entre les séquences, pour donner un poids relatif à la similarité entre les séquences comparées et diminuer l’information redondante pour les paires de séquences hautement similaires. La fonction objectif à être optimisée pour un alignement multiple (A) est donnée par :

$$AlignmentCost(A) = \sum_{i=2}^N \sum_{j=1}^{i-1} W_{i,j} COST(A_i, A_j) \quad (3)$$

où *COST* est le score d’alignement entre les séquences i et j et $W_{i,j}$ leur poids dans la MSA [279].

Une autre approche itérative pour l’alignement multiple de séquences qui utilise l’optimisation d’une fonction objectif, comme celle présentée dans l’équation 3, se base sur les algorithmes génétiques. Une implémentation populaire de cette méthode se reflète dans le logiciel SAGA¹⁷

17. SAGA : «*Sequence Alignment by Genetic Algorithm*»

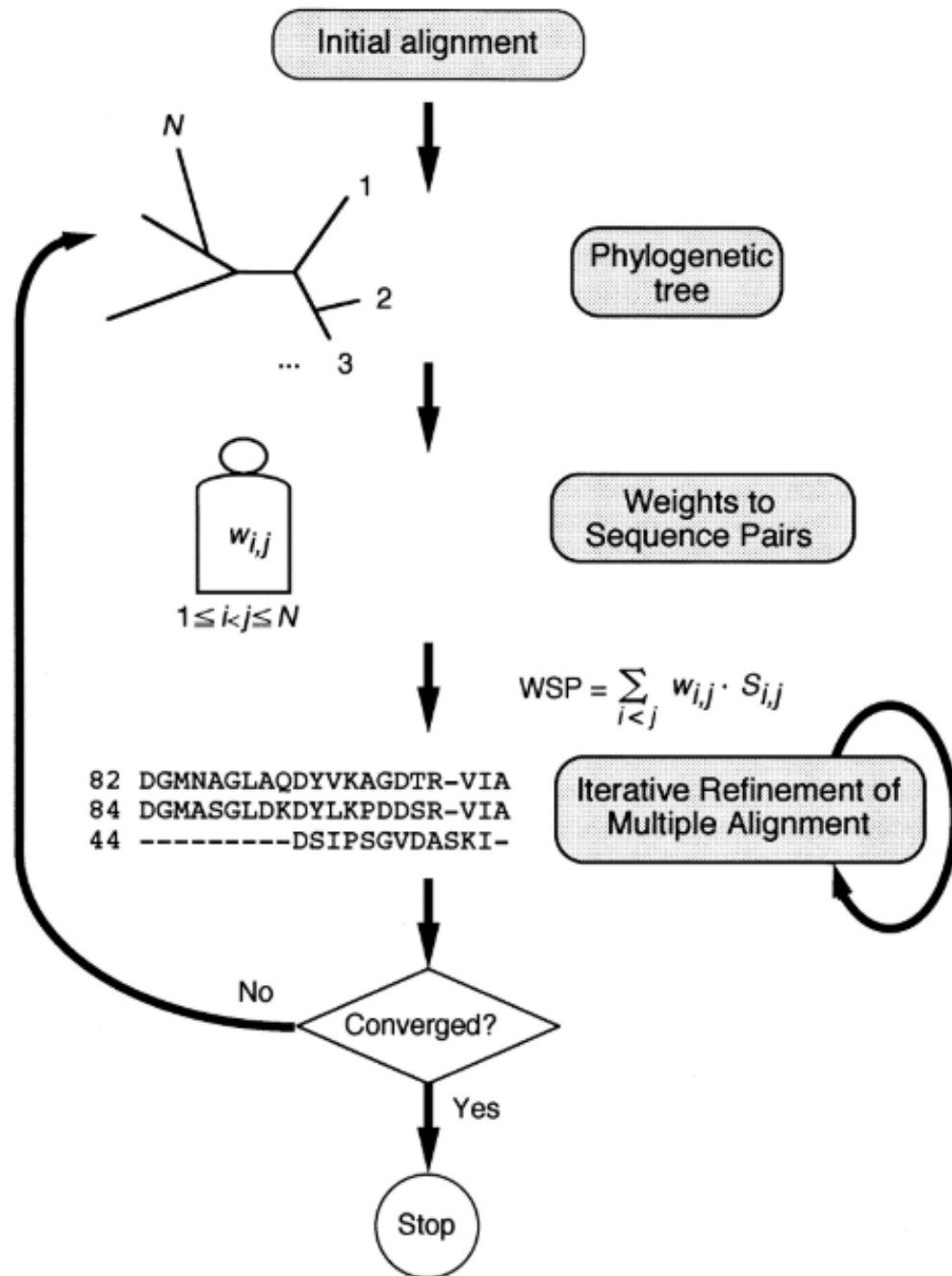


FIGURE I.14 – Illustration de l’algorithme de raffinement itératif pour l’alignement de séquences multiples. Figure issue de Gotoh et al. [147].

[279]. Brièvement, l’algorithme génétique implique l’utilisation d’une population de solutions qui évolue par sélection naturelle. Initialement, une première génération (G_0) est générée aléatoirement avec une taille de population constante. Pour passer d’une génération à une autre, des “alignements enfants” sont générés à partir des “alignements parents” par sélection na-

turelle selon une valeur sélective (*fitness*) qui se base sur la fonction objectif. La sélection naturelle se produit soit par *crossover* (un alignement mixte entre ceux des deux parents) ou par mutation (modification de l'alignement d'un seul parent). La procédure est répétée en ajustant les probabilités de sélection naturelle pour produire la prochaine génération d'alignements enfants et se terminera lorsqu'aucune amélioration ne sera observée sur plusieurs générations.

Finalement, les modèles de Markov caché (HMM) sont considérés des modèles probabilistes qui sont liés à l'alignement multiple pouvant aussi servir à l'identification de séquences [102]. En effet, un profil HMM permet de représenter dans un graphe acyclique les probabilités d'une brèche, d'un appariement ou d'un mésappariement pour chacune des positions d'un MSA. Tout comme dans les alignements progressifs, l'ordre dans lequel seront rajoutées les séquences pour l'alignement multiple peut avoir un impact sur la qualité du profil, cependant à chaque ajout d'une séquence, les autres séquences alignées seront revisitées pour assurer une meilleure qualité de l'alignement final. La figure I.15 illustre un profil HMM sous forme de graphe pour une courte séquence d'acides aminés. L'une des implémentations les plus

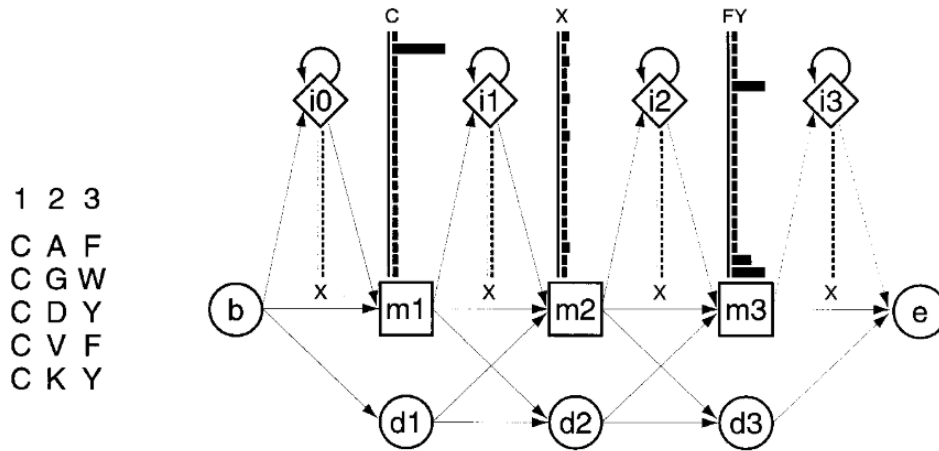


FIGURE I.15 – Modèle HMM d'un court alignement de cinq séquences avec trois positions. Les trois positions sont représentées par un état d'appariement (carrés *m*), un état de délétion (cercle *d*) ou un état d'insertion (losanges *i*). Chaque carré d'appariement possède une probabilité pour chacun des acides aminés possibles à cette position. Figure issue de Eddy et al. [102].

populaires à utiliser les profils HMM à des fins de comparaison de séquences est HMMER [102; 186; 264] et la base de données de profils la plus connue, Pfam [106]. En pratique, deux approches sont généralement utilisées pour produire l'alignement d'une séquence à un profil HMM, soit avec l'algorithme de Viterbi ou l'algorithme *forward* [101]. Les deux algorithmes ont une complexité algorithmique de $O(NM^2)$ pour le temps d'exécution et $O(NM)$ en espace, où N est la longueur de la séquence à identifier et M le nombre d'états dans le profil HMM [102].

I.3.3 Assemblage de génomes

Étant donné l'importance attribuée aux k -mers dans cette thèse, présentons leur utilité dans un autre aspect central en génomique, à savoir l'assemblage de génomes. L'assemblage de génomes consiste à partir des lectures d'ADN (souvent de petite longueur) produites par une expérience de séquençage (section I.2) et de les assembler en séquences d'ADN contiguës (*contigs*) représentatives des molécules d'ADN originales.

Pour la première génération de séquençage (sous-section I.2.1), comme avec le séquençage de Sanger, des algorithmes de consensus et de chevauchement étaient utilisés pour générer les assemblages. Deux assembleurs connus pour utiliser le consensus et le chevauchement sont Newbler de la compagnie 454 Life Sciences [254] et Celera Assembler par Denisov et al. [82]. Essentiellement, les lectures sont alignées pour produire des alignements multiples et ensuite un consensus est extrait de l'alignement multiple avec l'identification des bases les plus représentées à chaque position. L'identification d'allèles et de variants peut aussi être faite lorsqu'un consensus non unanime est identifié à une position donnée. La procédure s'avère possible à cause de la bonne qualité des lectures d'ADN de la méthode de Sanger, ainsi que de leur taille adéquate (≈ 800 paires de bases).

La deuxième génération de séquenceurs commerciaux d'ADN (sous-section I.2.2) est caractérisée par la diminution de la taille des lectures d'ADN, mais offre un débit et un volume de séquençage beaucoup plus important et pour une production finale en quantité de lectures plus élevée et plus abordable. Le lecteur est prié de se référer à la figure I.1 et au tableau I.1 pour la diminution du prix du séquençage en lien avec la parution des nouvelles technologies. La longueur limitée des lectures a poussé l'adoption de nouvelles techniques algorithmiques pour réaliser les assemblages. En effet, Pezner et collaborateurs introduisent en 2001 l'utilisation des k -mers (formellement l -tuples dans l'article) et le graphe de De Bruijn pour réaliser des assemblages *de novo* de génomes [304]. Leurs travaux attaquent le problème de l'assemblage *de novo* avec l'approche de graphe eulérien qui consiste à trouver des chemins dans le graphe de De Bruijn et qui mèneront ultimement à la création des *contigs*. Les graphes de De Bruijn deviendront alors le standard pour l'assemblage de courtes lectures des appareils de nouvelles générations de séquençage remplaçant ainsi les algorithmes de consensus et de chevauchement typique à la première génération de séquençages.

Plusieurs logiciels pour l'assemblage *de novo* utilisant les graphes de De Bruijn verront le jour avec différentes améliorations et variantes, dont EULER [305], Velvet [437], ABySS [361], Ray assembler [33] et SPAdes [24] pour ne nommer que ceux-là. L'un des désavantages des courtes lectures et de leur assemblage est la résolution des répétitions. Lorsqu'une répétition dans un génome (comme un gène répété) est de plus grande taille que la longueur des lectures, il devient difficile pour l'assembleur de connaître les voisins exacts de la répétition durant la création des *contigs*. L'un des meilleurs logiciels pour l'assemblage de génomes des bactéries

est SPAdes [24] qui utilise différentes tailles de k -mers pour la réalisation de l'assemblage. En effet, la variabilité de couverture des lectures sur l'ensemble d'un génome affecte l'assemblage et l'utilisation de k -mers plus courts dans les régions de faible couverture et une taille plus grande dans les régions de forte couverture améliorerait la qualité des assemblages. Pour cette raison, SPAdes utilise différentes longueurs de k -mers pour réaliser ses assemblages et reporte des résultats très intéressants [24]. Dans Magoc et al. [249], différents assembleurs sont évalués sur douze génomes de bactéries et SPAdes reporte les meilleurs résultats *ex aequo* avec MaSuRCA [443]. Au moment de l'écriture de cette thèse, le nombre de citations des deux articles penche en faveur de SPAdes le rendant ainsi, selon toute vraisemblance, le plus populaire pour l'assemblage *de novo* de génomes bactériens à partir de lectures d'ADN provenant des séquenceurs de deuxième génération.

La troisième génération de séquenceurs (sous-section I.2.3) se distingue de la deuxième génération de par une taille plus importante des lectures, malgré une moins bonne précision des lectures individuelles. Avec la venue de ces longues lectures, réapparaît aussi l'utilisation de l'assemblage avec chevauchement contrairement aux graphes de De Bruijn de la seconde génération. Un assembleur populaire pour les lectures produites par les technologies SMS de PacBio et d'Oxford Nanopore est Canu [210]. Canu est un successeur de l'assembleur Celera utilisé pour les lectures de la première génération de séquençages et qui utilise le «*overlap layout consensus*», aussi appelé consensus et chevauchement. À noter que les assemblages de la troisième génération avec leur longueur supérieure et les nouveaux algorithmes de chevauchement apporteront une nette amélioration sur les assemblages en termes de nombre de *contigs* couvrant la majeure partie des génomes. En effet, les courtes lectures de la deuxième génération de séquenceurs avaient comme désavantages de ne pas couvrir les répétitions dans les génomes bactériens, notamment des gènes de l'ARN ribosomal, et de mener à des assemblages incomplets avec plusieurs *contigs*, allant même jusqu'à une centaine. Il est donc maintenant possible avec les nouvelles technologies de séquençages d'obtenir un plasmide ou même un génome complet en un seul *contig*, évitant ainsi les efforts requis pour fermer les brèches dans les assemblages avec des PCR [216]. Puisque les courtes lectures sont généralement reconnues pour être de meilleure qualité que les longues lectures, une approche hybride qui mixe les courtes et les longues lectures offrira une séquence optimale en termes de qualité et de complétude.

I.3.4 Annotation de génomes

Avec la parution du séquençage, la prochaine étape logique était la caractérisation des gènes et de leur produit protéique allant jusqu'à leur fonctionnement au niveau cellulaire. Beaucoup des bactéries séquencées depuis l'aube du séquençage sont des pathogènes, comme le démontre les deux premiers génomes étant *Haemophilus influenzae* et *Mycoplasma genitalium* parût tous deux en 1995 [120; 126]. Dans un contexte de pathogénicité des bactéries, les éléments

génomiques sur lesquels on s'attarde souvent sont les gènes de résistances aux antibiotiques, les éléments mobiles qui servent à leur dissémination et les gènes de virulence. De ce fait, en plus des banques de données publiques (NCBI¹⁸, EBI¹⁹, DDBJ²⁰) beaucoup de bases de données spécialisées sur des gènes d'intérêts ont vu le jour pour les annotations de génomes bactériens.

Les formats de fichier d'annotations de génomes prennent habituellement l'une des formes suivantes : un fichier GenBank du NCBI, un fichier EMBL²¹ de l'EBI ou encore un fichier GFF²². Les formats de fichiers, tout juste mentionnés, sont respectivement présentés dans les figures annexes A5, A6 et A8. Les exemples de fichiers d'annotations présentent le plasmide pKp199-1 de la souche *Klebsiella pneumoniae subsp. pneumoniae* isolée dans la ville de Québec et publié en 2020 par Déraspe et al. [88]. La souche, isolée sur un patient en provenance de l'Inde, avait d'importance clinique principalement un gène de résistance aux antibiotiques carbapénèmes.

On remarque une ressemblance dans les fichiers d'annotations, disons plus classique (GenBank et EMBL), avec un code pour définir les types d'attributs de la séquence génomique avec la valeur qui suit sur différentes lignes du fichier, comme le code à deux lettres en début de ligne dans les fichiers EMBL et les attributs précédés d'une barre oblique dans les fichiers GenBank. Les fichiers GFF contiennent toute l'information d'un gène sur une seule ligne du fichier et utilisent comme séparateur le caractère `<TAB>` possiblement dû à sa popularité dans les fichiers généralement traités en bio-informatique.

Les logiciels d'annotations de génomes bactériens sont généralement des pipelines d'annotations, c'est-à-dire une suite d'outils qui prennent en entrées un format de données précis pour produire un résultat aussi dans un format bien précis qui sera ensuite pris en charge par l'outil suivant dans le pipeline. Ultimement, les annotations de génomes pourront être dans l'un des formats mentionnés précédemment, elles seront inspectées pour ses gènes d'intérêts et comparées à d'autres génomes pour des fins d'analyses.

Plusieurs pipelines existent donc pour l'annotation de génomes. Introduisons d'abord le pipeline du NCBI PGAP (*Prokaryotic Genome Annotation Pipeline*) qui sert d'ailleurs à introduire les nouveaux génomes séquencés dans NCBI RefSeq [287], l'une des bases de données les plus utilisées pour l'analyse de génomes bactériens. La figure I.16 présente le pipeline de PGAP qui est typique d'un pipeline d'annotation de génome de bactéries. Dans les procédures d'annotations, on retrouve habituellement l'identification de gènes codants à partir d'un génome assemblé (avec un *ORF-finder*²³), l'identification des gènes (ou protéines préalablement iden-

18. NCBI : *National Center for Biotechnology Information*

19. EBI : *European Bioinformatics Institute*

20. DDBJ : *DNA Data Bank of Japan*

21. EMBL : *European Molecular Biology Laboratory*

22. GFF : *General Feature Format*

23. ORF : *Open Reading Frame*

tifiées) en se basant sur des bases de données de gènes déjà caractérisées et la création de fichiers d'annotations pour leur soumission dans les bases de données publiques, comme celles du NCBI, de l'EBI et du DDBJ.

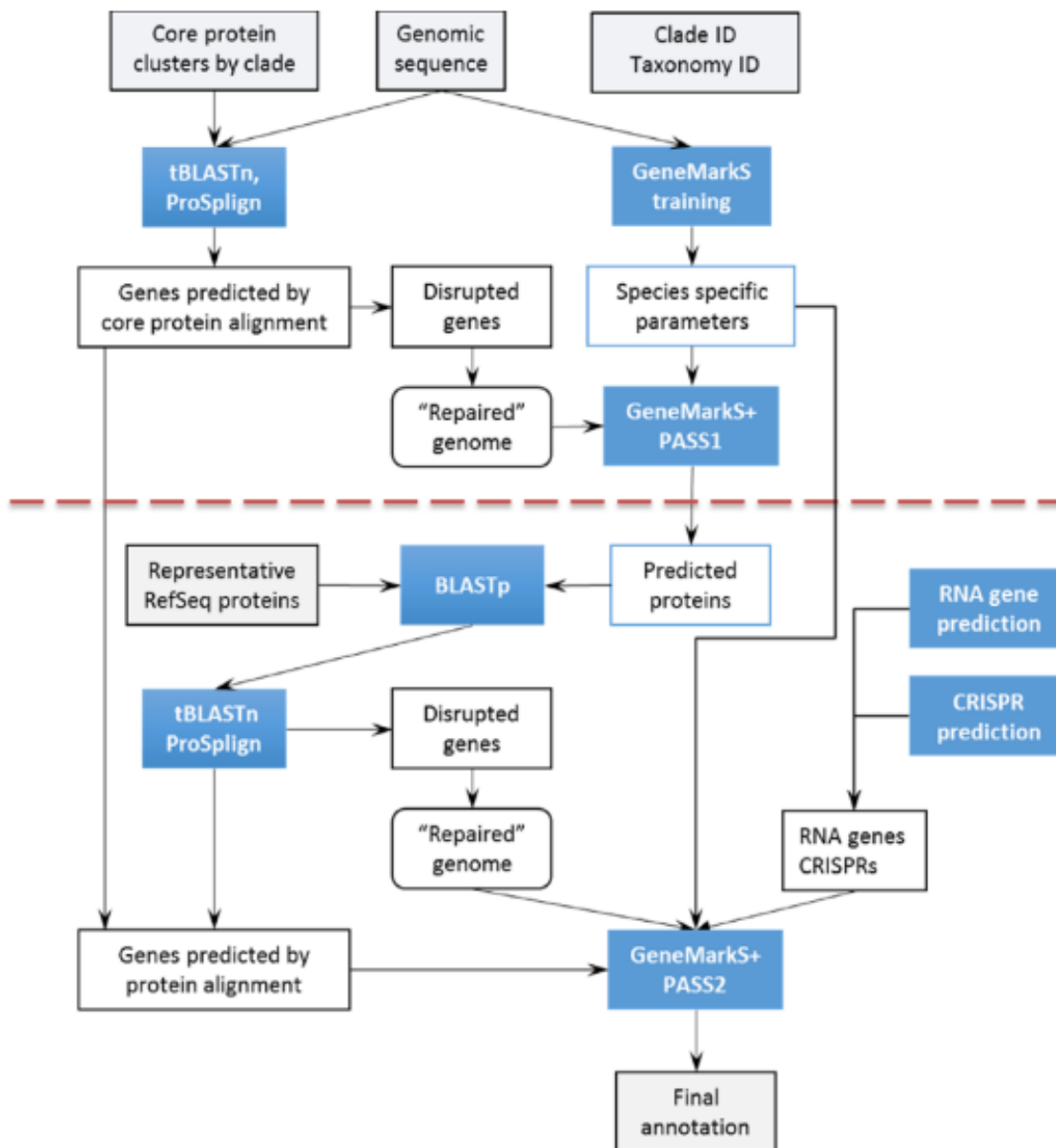


FIGURE I.16 – Pipeline d'annotation PGAP (*Prokaryotic Genome Annotation Pipeline*) de génomes bactériens du NCBI. Voir le texte pour plus de détails sur les différentes étapes du pipeline. Figure issue de Tatusova et al. [384].

Pour prendre particulièrement PGAP, une première passe est effectuée sur un assemblage en associant le génome à un clade taxonomique pour lequel un ensemble de protéines conservées est disponible pour faire l'identification. Le logiciel d'alignement tBLASTn [46] est utilisé pour

identifier les protéines spécifiques au clade directement à partir de l'ADN du génome étudié. Parallèlement, la prédiction de séquences codantes dans les assemblages est conduite avec GeneMarkS+ [28]. Les protéines identifiées seront comparées avec BLASTp à RefSeq pour leur identification. Le pipeline étant la couche logicielle qui orchestre le flux des données entre les différents outils, il s'occupera de résoudre les conflits ou duplicatas dans l'annotation du génome. Les ARNs, principalement de transfert et ribosomiques, de par leur particularité sont identifiés avec d'autres bases de données et différentes techniques de celles employées avec les gènes codants. Dans PGAP, ils utilisent la collection suivante : *NCBI RefSeq Targeted Loci* [382] qui inclut les gènes de l'ARN 16S et 23S. Pour les ARNs 5S et petites ARN non codant PGAP utilise *cmsearch*, alors qu'il utilise tRNAscan-SE [242] pour l'identification des ARN de transferts (tRNA).

Tout compte fait, les annotations sont grandement dépendantes de la qualité des jeux de données (séquences bien caractérisées) disponibles et qui sont utilisés comme référence dans l'identification de gènes et protéines dans les génomes bactériens. Le mémoire de maîtrise de l'auteur de cette thèse [84] présente d'ailleurs la construction d'un tel type de base de données, plus particulièrement sur la résistance aux antibiotiques, avec une démonstration de son utilisation dans une multitude d'analyses génomique et métagénomique [322; 321; 427; 425].

Beaucoup des analyses réalisées dans le cadre des études génomique et métagénomique tout juste citées, se sont basées à différents niveaux sur le *bacterial-annotator* ; un logiciel d'annotation de génome bactérien qui utilise le langage de programmation Ruby et la librairie bioruby [146]. Le logiciel *bacterial-annotator* fut implémenté par l'auteur de cette thèse est disponible en ligne (<https://github.com/zorino/bacterial-annotator>) sous la licence GNU GPL v3.0. De manière similaire à Prokka [351], il permet l'annotation de génomes à partir d'un génome ou d'une base de données de références et de construire des fichiers d'annotations de génomes bactériens, tels que vue aux annexes A5, A6 et A8. De plus, une plateforme Web, nommée Bacteriapps²⁴ est en développement pour faciliter l'utilisation des outils d'annotations et de comparaison de génomes implémentés dans le logiciel *bacterial-annotator*. Par ailleurs, un autre logiciel très pratique développé par l'auteur de cette thèse est *microbe-dbs*²⁵ qui permet de télécharger et de formater des bases de données disponibles publiquement et de les adapter aux analyses de génomique microbienne.

Le côté pratique des logiciels comme le *bacterial-annotator*, Prokka ou même YAMP²⁶ pour la métagénomique [402] est l'autonomie qu'ils procurent à l'utilisateur pour l'analyse de leurs données avant leur soumission à une base de données publique. En effet, les plateformes d'annotations comme celle du NCBI, RAST ou PATRIC [287; 20; 42; 410] offrent une plus grande

24. <https://bacteriapps.genome.ulaval.ca>

25. <https://github.com/zorino/microbe-dbs>

26. YAMP : *Yet Another Metagenomics Pipeline*

facilité d'exécution, en contrepartie une attente est requise pour obtenir les résultats, ces services étant publics et partagés par toute une communauté de scientifiques. Le tableau annexe A1 présente de manière non exhaustive plusieurs bases de données souvent utilisées dans l'annotation de génomes bactériens.

I.3.5 Comparaison de génomes par phylogénie et autres techniques

Au-delà de l'annotation de génomes, des méthodes existent pour comparer des génomes soit de manière *de novo* ou encore avec références. Une des méthodes de comparaison traditionnelles est la phylogénie avec la construction d'un arbre phylogénétique. En effet, un arbre phylogénétique permet de visualiser les relations évolutives entre les gènes ou génomes étudiés. Plusieurs méthodes de phylogénies existent et peuvent être généralement regroupées en deux catégories, soit celles basées seulement sur la distance entre les espèces (regroupement hiérarchique) et celles basées sur la différence de caractère (maximum de parcimonie, maximum de vraisemblance ou *maximum likelihood* (ML) et inférence bayésienne (IB)) [193].

Les méthodes de regroupement hiérarchique (*hierarchical clustering*), principalement UPGMA et NJ, sont souvent utilisées dans les algorithmes d'alignements multiples, tels que vus à la sous-section I.3.2. Cependant, elles sont généralement critiquées dans la construction de phylogénie étant données qu'elles ne tiennent pas compte d'un modèle évolutif pour les substitutions et indels²⁷ dans l'alignement des séquences sous-jacent à la phylogénie et aurait des performances plutôt médiocres pour les espèces éloignées [193]. Les méthodes prennent en entrée une matrice de distance qui représente la distance évolutive entre les génomes. Depuis cette matrice, des *clusters* sont progressivement créés avec une approche ascendante (*bottom-up*), soit en joignant les branches des génomes les plus similaires entre eux en premier. Le regroupement hiérarchique est donc dit agglomératif puisqu'au départ aucun *cluster* n'existe, à l'opposé les méthodes dites descendante ou *top-down* procèdent en divisant les observations en *clusters* à partir d'un seul groupement initial.

Maintenant pour ce qui est des méthodes phylogénétiques basées sur les différences de caractère, introduisons d'abord celle de parcimonie. L'utilisation de la méthode de maximum de parcimonie pour la phylogénie fut décrite par Farris et al. en 1970 et Fitch et al. en 1971 [110; 119]. Brièvement, la méthode calcule le nombre minimal de changements nécessaire, en termes de nucléotides pour l'ADN ou d'acides aminés pour les protéines, pour reconstruire la topologie des arbres visités. L'arbre possédant la topologie qui nécessite le moins de changement dans les séquences sous-jacentes sera celui considéré le plus parcimonieux. Une recherche exhaustive de toutes les possibilités de topologie d'arbre devient vite impossible plus le nombre d'espèces grandit, d'où la nécessité d'utiliser des heuristiques pour la recherche des topologies. L'un des désavantages de la méthode de maximum de parcimonie est l'absence d'un modèle d'évolution qui tient compte par exemple des différences entre les transitions (purine → purine et

27. Indel : Insertion ou délétion

pyrimidine \rightarrow pyrimidine) et transversions (purine \rightarrow pyrimidine et pyrimidine \rightarrow purine) pour l'ADN ou encore pour la substitution des acides aminés. De plus, la méthode serait particulièrement plus sensible au problème d'attraction des longues branches comparativement aux méthodes de maximum de vraisemblance (*maximum likelihood*) et d'inférence bayésienne (BI) [193].

Enfin, les méthodes ML et BI tiennent compte d'un modèle évolutif et de la probabilité que celui-ci soit bien représenté dans les données génomiques. Ces méthodes sont généralement considérées comme les mieux adaptés à la reconstruction d'une phylogénie représentative des réelles relations évolutives entre les spécimens étudiés. Les modèles supposent entre autres l'indépendance des différentes colonnes dans un alignement aussi nommé sites, et qui peuvent être des mutations, indels ou des bases et acides aminés conservés. À noter qu'avec l'indépendance des sites, il est possible de tenir compte de l'hétérogénéité du taux de substitutions par sites et peut être évalué avec un modèle comme GAMMA Γ [430] ou CAT («categories»), généralement implémenté dans les logiciels de phylogénie ML et BI. L'incorporation de ce paramètre d'hétérogénéité du taux de substitutions apporterait un gain significatif sur la reconstruction topologique des arbres phylogénétiques selon Yang et collaborateurs [431]. Les méthodes de vraisemblance peuvent donc être représentées comme le produit des probabilités observées à chacun des sites de l'alignement. Une probabilité à un site peut être défini comme une fonction $L(\theta)$ où θ inclut généralement les modèles de substitutions de nucléotides ou d'acides aminés avec leurs paramètres et la longueur des branches des arbres visitées dans l'espace de recherche de topologie des arbres.

La méthode de maximum de vraisemblance (ML), introduite par Felsenstein et al. en 1981 [113], consiste à trouver les paramètres qui maximisent la probabilité d'observer les données d'alignement de séquences. Les paramètres sont alors estimés pour maximiser une topologie qui sera représentative des données. Étant donné la complexité combinatoire pour estimer tous les paramètres, des approches non exhaustives sont généralement utilisées jusqu'à la convergence des données. Les logiciels les plus populaires pour la phylogénie ML sont PhyML [151; 149], FastTree [309; 310] et RaXML [374]. La première version de PhyML introduit l'utilisation de l'algorithme *hill-climbing* pour les itérations qui ajuste la topologie des arbres et la longueur des branches simultanément jusqu'à convergence [151]. Les paramètres du modèle, tels le ratio de transitions / transversions ou la fonction gamma pour les taux de substitutions par site, sont aussi évalués durant les itérations. La méthode *hill-climbing* est considérée plus rapide que l'optimisation stochastique et est souvent considérée comme suffisante lorsque la solution recherchée au problème se veut une simplification de la réalité. Tel est le cas pour une phylogénie, puisqu'il est impossible de connaître la séquence réelle des évènements qui ont menée aux substitutions observées dans les alignements. Dans la première version de PhyML, la recherche topologique du meilleur arbre est réalisée avec la méthode du plus proche voisin

(NNI²⁸) pour effectuer les permutations des sous-arbres durant les itérations. Il en est de même pour la première version de FastTree [309]. La version 2006 de RAxML, en plus d'améliorer les performances de l'outil avec la parallélisation MPI²⁹, implémentait un heuristique pour la recherche d'arbres basée sur l'élagage et la greffe de sous-arbres (SPR³⁰) [373]. Les plus récentes versions de PhyML (3.0) et de FastTree (2.1) ont toutes deux passés à l'utilisation du SPR pour la recherche topologique d'arbre pour contrer le fait que la méthode NNI a tendance à trouver des maximums locaux dans la fonction de maximum de vraisemblance et ainsi produire des effets indésirables comme l'attraction des longues branches [149; 310].

Outre la phylogénie, d'autres méthodes de comparaison de génomes existent, dont l'une d'entre elles est la moyenne d'identité nucléotidique (ANI³¹) [209]. La méthode ANI consiste à calculer la moyenne de pourcentage d'identité pour tous les gènes homologues partagés entre deux génomes. Le pourcentage d'identités se base sur l'alignement de séquences et peut être calculé avec un aligneur, tel BLAST ou même MUMMER. Il a été démontré qu'une ANI de 94% correspond au 70% d'hybridation ADN-ADN généralement utilisée pour la délimitation et la caractérisation des espèces bactériennes. L'ANI est d'ailleurs de plus en plus utilisée dans la comparaison à grande échelle de génomes bactériens et pour la délimitation des espèces de bactéries [180]. Plus de détails sur l'ANI en relation avec la taxonomie bactérienne seront présentés à la sous-section I.3.6.

Une autre méthode très rapide pour la comparaison génomique et ne nécessitant pas d'alignement est Mash, qui utilise la technique MinHash, une forme de *locality sensitive hashing*, utilisée pour la réduction de dimensionnalités [289]. Essentiellement, Mash utilise deux fonctions pour réaliser la comparaison de séquence : la première «*sketch*» permet de créer une signature d'une séquence ou collection de séquences basée sur les *k*-mers, alors que la deuxième «*dist*» permet d'évaluer la distance entre les *sketchs* de deux génomes. Les *sketchs* de Mash sont bâtis à partir des *s* (longueur du *sketch*) plus petites signatures générées avec une fonction de hachage sur l'ensemble des *k*-mers des séquences en entrées. Un *sketch*, malgré sa dimensionnalité significativement réduite par rapport aux séquences qu'il représente, permet quand même de produire des résultats de comparaisons qui corrèlent avec d'autres méthodes à base d'alignement, comme ANI. En effet, la fonction «*dist*» compare deux *sketchs* et retourne une estimation de l'index de Jaccard, une *P* – *value* et une distance Mash qui estime le taux de mutations entre les séquences. Mash peut aussi bien être utilisé pour la comparaison de séquences génomiques que métagénomiques.

Dans la même optique de comparaison génomique, nous avons développé un logiciel Ray Surveyor ne nécessitant pas d'alignement et qui permet d'évaluer la distance entre des séquences génomiques et métagénomiques. En effet, Ray Surveyor calcule le nombre total de *k*-mers

28. NNI : *Nearest Neighbor Interchanges*

29. MPI : *Message Passing Interface*

30. SPR : *Subtree Pruning and Regrafting*

31. ANI : *Average Nucleotide Identity*

partagés entre des séquences d'ADN et permet ainsi d'estimer la distance évolutive entre les génomes. Ray Surveyor est présenté au chapitre 1 avec plusieurs cas de démonstration en génomique bactérienne.

I.3.6 Taxonomie

À l'origine, les espèces bactériennes étaient décrites avec une batterie de tests biochimiques (source de carbone, source d'azote, coloration à Gram, etc.) et de leur morphologie cellulaire [216]. En se basant seulement sur un nombre limité de caractéristiques phénotypiques et non basé sur la génétique, fortes étaient les chances d'avoir une classification inconsistante des espèces. En effet, les techniques génétiques qui verront le jour permettront de révéler ces inconsistances et d'apporter des améliorations à la taxonomie des espèces bactériennes. Un exemple typique de classification erronée est celle des *Clostridium* qui sera revisitée par Yutin et al. [434]. Effectivement, le pathogène *Clostridium difficile*, étant donné son rapprochement évolutif avec les *Peptostreptococcaceae*, devrait maintenant être nommé *Peptoclostridium difficile*. Une des premières techniques génétiques pour classifier les espèces fut l'hybridation ADN-ADN. Elle devint ainsi la référence pour la délimitation des espèces avec un seuil de 70% de similarité ADN-ADN [411]. L'émergence des technologies de séquençage apporta une nouvelle méthodologie aux taxonomistes. En effet, étant donné l'ubiquité du gène de l'ARN ribosomal (ARNr) 16S chez les bactéries, son séquençage combiné à sa comparaison avec ceux des autres bactéries permet aussi la séparation des espèces. Le seuil de pourcentage d'identité pour séparer les espèces à partir de l'ARNr 16S généralement accepté est de 97% [372]. Donc, pour que deux génomes soient considérés de la même espèce, ils doivent partager au moins 97% d'identités entre leur ARNr 16S. L'utilisation de la séquence de l'ARNr 16S comme élément comparatif pour la délimitation des espèces se base sur plusieurs prémisses [206] :

1. le gène de l'ARNr 16s est ubiquitaire chez les bactéries ;
2. le gène de l'ARNr 16S est peu susceptible aux transferts horizontaux de gènes (THG) ;
3. sa longueur de 1500 bases est adéquate pour l'analyse comparative ;
4. sa séquence est assez conservée pour pouvoir comparer des espèces éloignées ;
5. certaines régions du gène sont parfaitement conservées permettant la conception d'amorces PCR ;
6. certaines régions du gène sont assez variables pour permettre la délimitation des espèces similaires.

Il existe quand même certaines limitations à l'analyse taxonomique à partir du gène de l'ARN 16S, comme le fait que plusieurs copies différentes du gène existent souvent au sein d'un même génome [185]. Avec l'arrivée des nouvelles générations de séquençage et leur coût très abordable suivi des avancées méthodologiques en bio-informatique, il est maintenant plus facile de réaliser la classification taxonomique des espèces à partir des génomes complets. En effet, «*The Genome Taxonomy Database*» (GTD) utilise l'ANI avec un seuil de 95% d'identités

pour déterminer les espèces des bactéries et archéobactéries au sein de leur base de données [185]. Tel que mentionné précédemment, Konstantinidis et al. avait démontré qu'un seuil de 94% d'identités corrélait bien avec la classification des espèces bactériennes au moment de la publication de leurs travaux [209]. D'autres méthodes pour la classification ont aussi été proposées, comme l'analyse de séquence multilocus (MLSA ³²) qui inclut plusieurs gènes ménagers des génomes pour la comparaison, au lieu de se limiter à un seul comme l'ARN 16S [346].

C'est dans le journal *International journal of systematic and evolutionary microbiology* où est publiée la caractérisation de nouveaux taxa bactériens (espèce, genre, famille ou autre rang). La figure I.17 illustre la procédure acceptée par le journal pour la caractérisation d'une nouvelle espèce bactérienne [62]. Le seuil d'identités pour le gène de l'ARN 16S est de 98.7%.

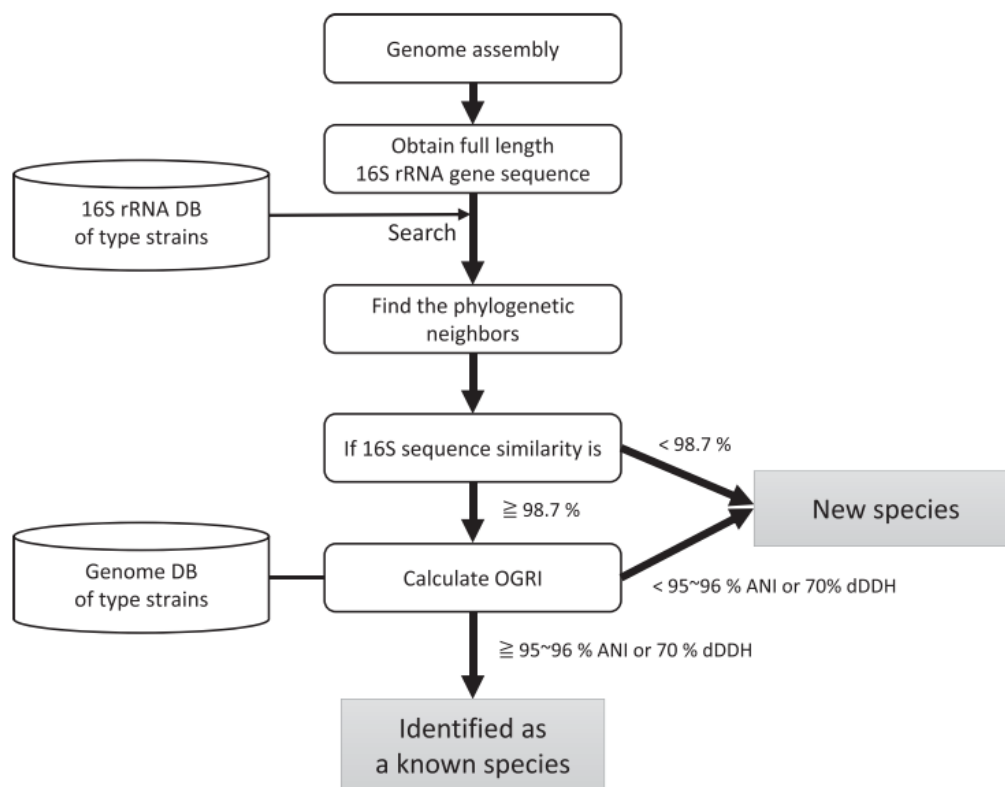


FIGURE I.17 – Diagramme pour l'identification de nouvelles espèces bactériennes. Figure tirée de Chun et al. [62]

Si la similarité surpasse ce seuil, alors il est nécessaire d'avoir un *overall genome related index* (OGRI) qui respecte l'un des critères suivants : soit un ANI < 95% ou encore une hybridation ADN-ADN < 70% pour que le génome soit considéré comme appartenant à nouvelle espèce. Étant donné que la résolution OGRI ne dépasse généralement pas le niveau taxonomique

32. MLSA : *Multilocus Sequence Analyses*

à l'espèce, pour les rangs plus élevés (genre, famille, etc.), une phylogénie avec des gènes conservés au-delà de l'espèce est nécessaire en plus de l'analyse du gène de l'ARN 16S.

Il est important de noter que la qualité de la classification taxonomique dans les bases de données de bactéries est primordiale aux analyses métagénomiques puisque l'une des principales analyses des métagénomomes est généralement la quantification des abondances des taxa bactériens qui s'y retrouvent. La prochaine section traitera plus en profondeur de la taxonomie dans les analyses métagénomiques.

I.3.7 Analyses prédictives en génomique

L'apprentissage automatique (*machine learning*) est une discipline de plus en plus utilisée en bio-informatique. En effet, pour réaliser des analyses prédictives, l'apprentissage automatique supervisé est souvent utilisé puisqu'il permet d'entraîner des algorithmes avec des jeux de données étiquetés³³ et d'en faire des prédicteurs avec des garanties statistiques. Plusieurs champs de recherches en bio-informatique et en génomique se servent donc de l'apprentissage automatique pour tenter d'améliorer l'aspect prédictif de leur analyse. On dénote entre autres chez les eucaryotes l'identification des sites d'initiation de la transcription [286], des sites d'épissage alternatif dans l'ARN codant [81] ou encore du positionnement des nucléosomes [352].

Chez les procaryotes, l'un des enjeux majeurs pour la santé humaine est leur résistance grandissante face aux antibiotiques utilisés dans le traitement des infections causées par les pathogènes. Il va sans dire que le phénomène a été profondément étudié et que beaucoup d'expérimentations et de logiciels basés sur l'apprentissage automatique ont été réalisés pour approfondir nos connaissances sur la génomique des bactéries en lien avec leur résistance aux antibiotiques.

Par exemple, Niehaus et collaborateurs ont analysé plus de 500 mutations ponctuelles chez *Mycobacterium tuberculosis* pour prédire leur résistance aux antibiotiques [276]. Les algorithmes testés étaient la régression logistique, le *support vector machine* (SVM) et une combinaison des deux avec la méthode d'association directe. Brièvement, la régression logistique (RL) est une forme de classification linéaire où des poids w sont assignés à chacune des caractéristiques dans un ensemble de données en vue de prédire un résultat qui concorde avec les classes étudiées (0 ou 1 dans le cas d'une régression binaire). La fonction utilisée avec cette méthode est la fonction logistique ou fonction sigmoïde (voir figure annexe A7) qui avec un seuil donné permet de classer un exemple comme étant positif ou négatif [44]. L'apprentissage se fait en optimisant le maximum de vraisemblance (*maximum likelihood*) depuis les probabilités données à chacune caractéristique. Étant donné qu'il n'existe pas de solution analytique au problème de régression logistique, l'apprentissage doit se faire avec une méthode d'optimisation, tel que la

33. Les jeux de données sont étiquetés pour un phénotype étudié.

descente de gradient. L'algorithme du SVM, quant à lui, procède essentiellement en projetant les caractéristiques dans un espace de haute dimensionnalité à l'aide d'un noyau mathématique (*kernel*) et tente de trouver un hyperplan pour séparer les classes [44]. Des résultats très intéressants furent donc obtenus par Niehaus et collaborateurs avec ces méthodes, particulièrement pour la prédiction de la résistance à l'antibiotique isoniazide avec une exactitude de 93%, autant avec la RL que le SVM.

Une autre étude ayant utilisé l'apprentissage automatique pour la prédiction de la résistance aux antibiotiques est celle de Pesesky et collaborateurs, qui ont fait l'exercice chez les *Enterobacteriaceae* [303]. Leurs expérimentations ont aussi utilisé la régression logistique avec comme caractéristiques la présence et l'absence de gènes de résistances aux antibiotiques identifiés au sein des génomes à partir de la base de données Resfams [137]. Ils ont aussi comparé les résultats entre l'apprentissage automatique et une approche axée seulement sur les règles de connaissances depuis les annotations des gènes de résistances dans la base de données. Les douze antibiotiques testés par les deux approches ont générés des résultats très similaires avec un léger avantage pour l'approche de régression logistique. En effet, la RL était en moyenne 90.3% en accord avec les tests de résistances *in vitro* contre 89.0% pour l'approche basée sur les règles de connaissances. Néanmoins, il est important de mentionner que les résultats sont intrinsèquement liés de par la nature de l'expérimentation qui utilisait seulement le contenu en gènes de résistances des génomes depuis l'annotation avec les bases de données pour les deux méthodes évaluées.

Les plateformes PATRIC et RAST ont chacune réalisé leur projet d'envergure concernant la prédiction de la résistance aux antibiotiques à l'aide de l'apprentissage automatique [78]. L'étude consistait à évaluer et développer des prédicteurs pour la résistance aux carbapénèmes chez *Acinetobacter baumannii*, la résistance à la méthicilline chez *Staphylococcus aureus*, la résistance aux bêta-lactame et cotrimoxazole chez *Streptococcus pneumoniae* et la résistance à de multiples antibiotiques chez *Mycobacterium tuberculosis*. L'algorithme utilisé dans leur expérimentation est AdaBoost (*Adaptive Boosting*) [127] et les caractéristiques utilisées pour l'apprentissage sont la présence ou l'absence des k -mers issus des *contigs* assemblés de tous les génomes analysés. AdaBoost est un algorithme typique de *boosting* qui obtient de bons résultats en combinant un ensemble de classificateurs faibles sur lesquels des poids sont attribués pour en faire un modèle final qui surpasse l'ensemble des modèles individuels. L'algorithme procède de manière itérative en identifiant les k -mers qui classifient le mieux les données et en recentrant l'apprentissage sur les éléments mal classifiés dans les rondes précédentes, jusqu'à l'atteinte d'un minimum global. Les résultats reportés par Davis et collaborateurs présentent différents niveaux d'exactitude selon l'espèce et l'antibiotique analysés, mais sont généralement assez bons pour conclure de manière générale qu'il est souvent possible d'expliquer la résistance aux antibiotiques depuis les génotypes des espèces étudiées à l'aide de l'apprentissage automatique. Dans les limitations à l'approche, on compte l'expression génique où le phéno-

type d'intérêt n'est pas directement visible dans la séquence d'ADN, ou encore une quantité limitée d'exemples de génomes positifs ou négatifs reliés au phénotype.

L'auteur de cette thèse a aussi participé à une étude d'apprentissage automatique en lien avec la résistance aux antibiotiques. En effet, l'étude de Drouin et collaborateurs introduisait une implémentation du *set covering machine* (SCM) [252], nommé *Kover*³⁴ et de son utilisation pour la prédiction de la résistance à 17 antibiotiques chez quatre espèces bactériennes (*Clostridium difficile*, *Mycobacterium tuberculosis*, *Pseudomonas aeruginosa* et *Streptococcus pneumoniae*) [98]. De manière similaire à l'étude de Davis et al., l'absence et la présence de k -mers sont utilisées pour l'entraînement du SCM. L'expérimentation utilisait d'ailleurs Ray Surveyor (chapitre 1) pour l'identification des k -mers au sein des génomes analysés. Essentiellement, le SCM utilise un algorithme glouton (*greedy algorithm*) pour produire un modèle de règles parcimonieuses, soit de conjonctions ou de disjonctions. Le résultat pour expliquer un phénotype, telle la résistance aux antibiotiques, est donc la présence ou l'absence de k -mers combinée soit avec l'opérateur logique "ET" (présence/absence des k -mers k_1 et k_2 et k_3) ou l'opérateur logique "OU" (présence/absence des k -mers k_1 ou k_2 ou k_3). La qualité des prédictions du SCM, telle que reportée dans l'étude, se compare à d'autres algorithmes, tels que les arbres de décisions ou le SVM, mais donne des résultats plus parcimonieux sur l'ensemble des données analysées.

En résumé, les méthodes d'apprentissage automatique peuvent s'avérer très utiles pour élucider, de manière *de novo*, les génotypes expliquant des phénotypes d'intérêts en génomique microbienne.

34. <https://github.com/aldro61/kover/>

I.4 Analyses métagénomique des bactéries

La majorité des techniques d'analyse de génomes de bactéries, présentées à la section précédente (I.3), sont aussi utilisées dans les analyses de métagénomies. Par exemple, la préparation des lectures d'ADN, l'identification, l'assemblage et l'annotation de séquences biologiques sont des procédures qui sont aussi utiles en métagénomique. Par contre pour certaines analyses, des logiciels spécialisés pour la métagénomique introduiront des modifications pour s'adapter aux particularités des métagénomies. En effet, un métagénome se doit d'être analysé comme un écosystème de microbes, ce qui diffère de la comparaison de génomes, généralement d'une même espèce où la similarité entre les spécimens est très grande. De plus, la variabilité des métagénomies entre les individus est tel que Franzosa et collaborateurs ont pu démontrer qu'il était possible d'obtenir une signature unique pour chacun des individus dans un groupe de 100 [124]. Similairement, une étude de 2015, dont l'auteur de cette thèse a participé, a démontré que la variabilité interindividuelle des espèces bactériennes du microbiote intestinal était plus grande que l'impact de l'antibiotique sur chacun des individus [322]. Malgré la grande divergence des métagénomies, il demeure souvent possible d'extraire des signaux sur les variations métagénomiques qui se produisent chez un individu exposé à un facteur externe, comme la prise d'un médicament, une diète spécifique ou une condition physiologique particulière. Un survol des différentes conditions généralement étudiées en lien avec le métagénome intestinal humain sera présenté à la prochaine section I.5. La présente section traite des analyses typiques en métagénomique ainsi que des logiciels généralement utilisés pour mener à bien le traitement des données de séquençage jusqu'à leur interprétation en lien avec une intervention ou un phénotype.

La figure I.18 présente un *workflow* typique des analyses en métagénomique, tel que présenté dans l'article de Knight et al. sur les meilleures pratiques conseillées en métagénomique [207]. Le deuxième niveau du diagramme (carré en rose) introduit trois types d'analyses, soit le profilage des communautés de microbes ainsi que leur profilage fonctionnel et en temps réels. Les prochaines sous-sections I.4.1 et I.4.2 introduiront le profilage taxonomique et le profilage fonctionnel, avec un accent particulier sur les standards généralement acceptés par la communauté de chercheurs pour les analyses bio-informatiques en métagénomique.

I.4.1 Profilage taxonomique

L'analyse de profilage taxonomique dépendra de la méthode de séquençage utilisée. Introduisons les logiciels et analyses bio-informatiques pour le profilage taxonomique, d'abord pour les approches par amplicon et ensuite pour celle par *shotgun*.

La première étape après avoir reçu les lectures d'ADN d'un séquenceur, autant pour les amplicons que le *shotgun*, est leur préparation. La préparation des lectures d'ADN a déjà été présentée à la sous-section I.3.1 de la section sur les analyses génomiques. Pour la méthode par

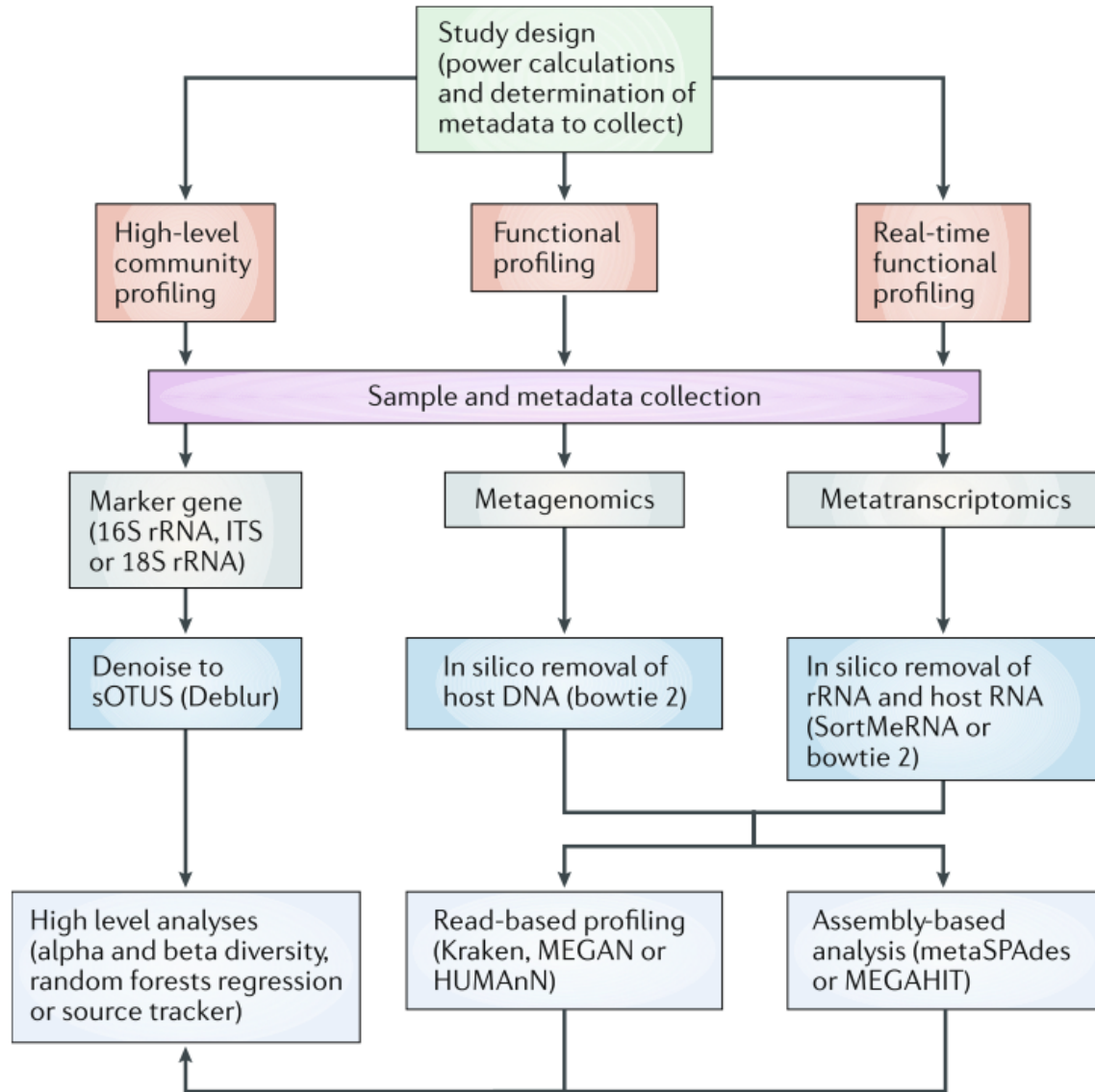


FIGURE I.18 – Flux de travaux conseillé par Knight et al. dans son article sur les meilleures pratiques pour l’analyse de métagénomiques basée sur le gène de l’ARN 16S, de métagénomique par *shotgun* et de métatranscriptomique [207].

amplicon, les séquences des lectures seront regroupées par unité taxonomique opérationnelle (OTU³⁵) généralement avec un seuil d’identités avoisinant 97%, tel qu’utilisé en taxonomie pour la définition de nouvelles espèces bactériennes (voir I.3.6) [207]. Par ailleurs, il est généralement conseillé d’utiliser des méthodes d’oligotypage (*oligotyping*) qui inclut de l’information supplémentaire sur les séquences d’ARN 16S, versus un simple seuil de 97% d’identités, dans le but d’améliorer la précision d’assignation des lectures aux OTUs, nommé *sub-OTUs* [108]. Des logiciels comme DADA2 [45] ou Deblur [10] permettent l’analyse d’oligotypage et sont disponibles dans la plateforme d’analyse métagénomique QIIME2 [36]. Ensuite, une base de

35. OTU : *Operational Taxonomic Unit*

données de références pour l'association taxonomique entre les OTUs ou sub-OTUs et les taxa bactériens est utilisée pour le profilage. L'une des trois bases de données suivantes est généralement utilisée pour l'association taxonomique avec les OTUs de l'ARN 16S : SILVA [433], RDP [69] ou Greengenes [91].

Pour la métagénomique par *shotgun*, le profilage taxonomique peut se faire à l'aide de méthode par *k*-mers, comme avec les logiciels kraken2 [420] ou CLARK-S [292], ou encore par l'identification de marqueurs avec des aligneurs de lectures d'ADN, telle qu'utilisée dans le logiciel MetaPhlAn2 [390] ou mOTUs [376].

Kraken, par exemple, utilise les *k*-mers et leur association au dernier ancêtre commun (LCA³⁶) dans la taxonomie pour permettre la quantification des taxa dans les métagénomés analysés. Cette méthode se veut plutôt gourmande en mémoire pour les analyses bio-informatiques, puisque tous les *k*-mers sont conservés en mémoire. Une amélioration à la méthode introduite dans Kraken2, est justement la diminution de mémoire requise pour effectuer les analyses de par l'utilisation d'une table de hachage compact qui utilise des *minimizers* au lieu de la totalité des *k*-mers pour ses entrées clé-valeur. Le lecteur est prié de se référer à la partie sur les compteurs de *k*-mers de la section I.3.2 pour plus de détails sur les *minimizers*. La base de données généralement utilisée avec Kraken2 est MiniKraken qui inclut des séquences de bactéries, archéobactéries et virus provenant de NCBI Refseq. La quantification des abondances des taxa dans les métagénomés peut ensuite être évalué avec Bracken qui est compatible avec les versions 1 et 2 de Kraken [245]. Bracken («*Bayesian Reestimation of Abundance after Classification with KrakEN*») agit en redistribuant l'assignation des lectures dans l'arbre taxonomique. Par exemple, si une lecture d'ADN est assignée au niveau taxonomique du genre, Bracken redistribuera les abondances au rang taxonomique inférieur, soit à l'espèce et aux niveaux supérieurs (famille, ordre, etc.) Ainsi, les abondances relatives tiendront compte, de manière probabiliste, de tous les rangs taxonomiques même si l'assignation se fait à un rang particulier.

Contrairement à Kraken, le logiciel MetaPhlAn2 («*Metagenomic Phylogenetic Analysis*») utilise l'alignement de séquences avec l'aligneur de courte lecture d'ADN : Bowtie2 [219; 390]. La quantification des abondances relatives se fait donc avec le compte des identifications des lectures d'ADN avec les marqueurs de la BD de références. MetaPhlAn2 inclut plus de 1 million de marqueurs spécifiques à différents clades et provenant d'au moins 7,500 espèces dans sa base de données. Dans leur étude, MetaPhlAn2 reporte des résultats supérieurs à kraken et mOTUs, générant moins de faux positifs et de faux négatifs durant l'évaluation des abondances taxonomiques sur 24 métagénomés synthétiques.

Cependant, dans l'étude de Ye et collaborateurs un étalonnage a été réalisé avec plusieurs logiciels de classification taxonomique et les méthodes par identification de marqueurs ne

36. LCA : *Last Common Ancestor*

semblaient pas mieux performer que celles par k -mers [432]. En effet, les classificateurs les plus performants sur les métagénomés synthétiques sont basés sur les longs k -mers d'ADN ($k > 30$ nt), comme kraken et taxMaps. Par contre, le facteur semblant le plus impacter les résultats de classification serait le contenu de la base de données de référence plus que la méthode d'annotation même. Pour ce qui est de la performance des logiciels, la majorité des méthodes requièrent une quantité importante de mémoire et de temps de calcul pour réaliser la classification taxonomique d'un métagénome. De plus dans l'étalonnage, les méthodes qui procédaient avec l'identification de protéines semblaient être les plus lentes, possiblement dû au temps requis pour la translation des six cadres de lecture de l'ADN. Dans cette thèse, le chapitre 2 présente un logiciel performant pour l'identification de protéines à partir de k -mers d'acides aminés (kAAmer) ainsi que son utilisation en métagénomique dans le chapitre 3.

Pour comparer le contenu taxonomique entre différents métagénomés, deux indices de diversité sont généralement utilisés : la diversité alpha et la diversité bêta. La diversité alpha calcule la richesse en espèce (ou taxa) au sein d'un métagénome avec des indices tel que Chao1 [53] ou Shannon [369]. Ensuite, la moyenne de diversité entre les métagénomés peut être comparée avec ces indices et est souvent interprétée en relation avec une condition étudiée. À noter que la diversité alpha est calculée au sein d'un seul métagénome. À l'opposé, l'index de diversité bêta calcul la dissimilarité entre les métagénomés et produit une matrice de distance entre chacune des paires des métagénomés étudiés. Les métriques quantitatives comme Bray-Curtis [40] ou Unifrac [243] sont généralement utilisées pour calculer la dissimilarité entre les métagénomés en tenant compte de l'abondance des taxa, alors que les métriques qualitatives comme Jaccard tiennent seulement compte de la présence et de l'absence des taxa. Ensuite, des tests non paramétriques comme PERMANOVA [13] ou ANOSIM [65] sont réalisés pour les analyses multivariées, question d'évaluer les différences entre les métagénomés des groupes étudiés. Pour la visualisation des données de bêta diversité, des techniques de réduction de dimensionnalité comme l'analyse en coordonnées principales (PCoA³⁷) ou l'analyse en composantes principales (PCA³⁸) sont nécessaire étant donné la complexité et l'imposante dimensionnalité des matrices de distances de bêta diversité.

Il existe plusieurs défis en ce qui concerne l'identification des taxa microbiens pour expliquer une condition entre deux groupes de métagénomés. Parmi ces défis, on compte notamment la haute dimensionnalité des données, leur état parcimonieux et compositionnel. La compositionnalité vient du fait que les abondances sont relatives, dû au limitation du séquençage, et que l'augmentation de la proportion d'un taxon implique nécessairement la diminution des autres taxa pour que le tout somme à 1. Ce problème a fait l'objet d'études statistiques et des méthodes de normalisation des données avec des transformations basées sur le *log-ratio* ont été développées pour contrer le problème de compositionnalité. L'une de ces transformations est

37. PCoA : *Principal Coordinate Analysis*

38. PCA : *Principal Component Analysis*

le *centered log-ratio* (CLR) introduit par Aitchison en 1982 [3]. Une transformation de ratio capture la dépendance dans les données et demeure identique autant pour les comptes bruts que pour les proportions des abondances taxonomiques. Le logarithme de ce ratio (log-ratio) rend les données symétriques et linéairement dépendantes [299]. Une propriété intéressante de la transformation de log-ratio est qu'elle projette les données dans un espace euclidien et que les méthodes statistiques adaptées aux nombres réels, comme les analyses multivariées, peuvent s'appliquer [140]. À titre d'exemple, les logiciels SparCC [128] et SPieCeasi [215] tiennent compte de la compositionnalité des données pour évaluer la corrélation entre les taxa au sein des métagénomés. Néanmoins, trouver une approche optimale et générique pour la compositionnalité est toujours une question de recherche pertinente [140].

I.4.2 Profilage fonctionnel

Le profilage fonctionnel des métagénomés, c'est-à-dire l'identification de la capacité des microorganismes à métaboliser différentes molécules, dépend grandement de la qualité des annotations des génomes déjà séquencés et caractérisés. L'organisation (*curation*) des connaissances sur les fonctions encodées par les génomes et leur publication dans des bases de données d'annotations sont cruciales à tout projet en génomique et métagénomique microbienne. Des projets tels que ENZYME database [21], Gene Ontology [18] et COG [381] sont souvent employés pour annoter les gènes et leurs fonctions alors que des projets tels que KEGG [285] et BioCyc [196] sont souvent employés pour l'identification des sentiers métaboliques.

Pour les analyses métagénomiques basées sur l'ARN 16S, l'identification directe des gènes impliqués dans les fonctions métaboliques des microbes n'est pas possible étant donné la nature du séquençage. Il est cependant possible d'extrapoler pour estimer la capacité fonctionnelle d'un métagénomés depuis le contenu en espèces microbiennes identifiées avec le séquençage de marqueurs. Un logiciel adapté pour réaliser de tels analyses est PICRUSt [217]. L'algorithme de PICRUSt procède en deux étapes. La première étape sert à inférer le contenu en gènes pour chaque espèce dans une phylogénie de référence. Les annotations des génomes références des bactéries et archéobactéries proviennent de la base de données IMG [255] et peuvent accommoder différentes annotations fonctionnelles comme celles des bases de données précédemment mentionnées. Dans leur article, Langille et collaborateurs font d'ailleurs la démonstration avec des données de KEGG et COG. La deuxième étape s'occupe d'inférer la capacité fonctionnelle des métagénomés en utilisant la phylogénie de référence et les abondances relatives provenant de l'identification du gène de l'ARN 16S dans les métagénomés. L'utilisateur doit fournir une table d'OTUs avec les identifiants de la base de données Greengenes, généralement produite avec une solution comme QIIME2. PICRUSt s'occupe ensuite de produire un tableau d'annotations avec les comptes de familles de gènes selon les annotations réalisées à la première étape. Les résultats seraient même ensuite directement comparables avec ceux obtenus depuis des expériences de métagénomique par *shotgun* et analysés avec des logiciels comme MG-RAST [201]

et HUMAnN2 [125].

Les analyses de métagénomique par *shotgun* permettent d'identifier directement les gènes impliqués dans les capacités fonctionnelles des microorganismes. Un exemple de logiciel pour ce type d'analyse est MG-RAST, un serveur d'analyse de métagénomiques disponible gratuitement en ligne et hébergé par l'université de Chicago et le laboratoire national Argonne [201]. Le pipeline d'annotation procède de manière similaire aux analyses typiques de séquences provenant des appareils de séquençage de nouvelles générations. D'abord, les lectures d'ADN sont préparées tel que vue à la section I.3.1 et il est aussi possible de supprimer les lectures de l'hôte (humain, souris, vache, etc.) des données, une étape qui est réalisée avec l'aligneur Bowtie. Par la suite, au lieu de faire l'identification des gènes d'intérêts directement dans les lectures d'ADN avec un aligneur, les fragments de gènes codant pour les protéines sont identifiés avec FragGeneScan [325], pour alléger l'intensité du calcul informatique nécessaire aux alignements des lectures. FragGeneScan utilise un modèle de Markov caché combinant des modèles d'erreurs de séquençage et d'utilisation de codon pour faire l'identification des séquences protéiques. Les protéines identifiées avec FragGeneScan qui partagent une identité d'au moins 90% en acides aminés sont regroupées avec le logiciel UCLUST [104] pour alléger leur identification. L'identification des protéines dans MG-RAST utilise une version modifiée de BLAT [202] (sBLAT) avec une base de données de protéines M5nr [414] qui inclut des annotations de plusieurs autres sources de données, tels que GO [18], KEGG [190] et COG [381]. MG-RAST calcule aussi les abondances taxonomiques qui sont projetées sur la taxonomie du NCBI [111].

Un autre logiciel très répandu pour l'analyse métagénomique par *shotgun* est HUMAnN2 [125]. HUMAnN2 emploie une stratégie de recherche à plusieurs niveaux (*tiered search*) que les auteurs revendiquent être plus performante et efficace qu'une simple identification direct des protéines dans les lectures d'ADN par traduction des cadres de lectures. La première étape du pipeline de HUMAnN2 est l'identification rapide de l'appartenance des lectures d'ADN aux espèces microbiennes avec MetaPhlan2 qui fut présenté dans la sous-section I.4.1 sur le profilage taxonomique. Une base de données pangénomique spécifique aux espèces identifiées par MetaPhlan2 est alors créée pour les prochaines étapes de l'analyse. La deuxième étape utilise donc la base de données pangénomique pour réaliser un alignement rapide des lectures du métagénome sur celle-ci. Une bonne partie des lectures se voit donc assigner un gène et une espèce depuis les bases de données de HUMAnN2. La troisième et dernière étape consiste à faire l'identification des lectures non assignées par la deuxième étape avec la base de données UniRef (UniRef90 ou UniRef50) [377]. Le mappage des lectures d'ADN aux gènes sert ensuite à estimer les abondances taxonomiques et fonctionnelles avec les différentes annotations incluses dans les bases de données de HUMAnN2, tels que GO [18], COG [381] et MetaCyc [195].

Certains pipelines métagénomiques, pour le séquençage par *shotgun*, utilisent l'assemblage des métagénomiques pour faire l'annotation des séquences et subséquemment le profilage fonctionnel

[378; 64]. L'un des avantages d'utiliser un assemblage pour l'identification des gènes au sein d'un métagénome, est d'obtenir le contexte dans lequel ceux-ci se retrouvent. Effectivement, lorsque les *contigs* assemblés sont suffisamment longs, on peut obtenir des parties de génomes et connaître les gènes voisins aux gènes d'intérêt. À titre d'exemple, les gènes de résistances aux antibiotiques sont souvent hébergés sur des éléments mobiles qui favorisent leur dissémination et cette information est seulement observable avec des *contigs* de taille suffisante. L'assemblage de métagénomes comporte plusieurs défis comparativement à l'assemblage de génome [19]. Tout d'abord, lors du séquençage d'un métagénome on ne connaît pas préalablement le nombre d'espèces séquencées et leur abondance. Pour le séquençage d'un seul génome, on connaît habituellement son origine et sa taille attendue, ce qui permet d'estimer la couverture du séquençage, une information qui peut s'avérer utile durant l'assemblage pour écarter les lectures issues d'erreurs de séquençages. De plus, dans un métagénome le chevauchement des ensembles de k -mers, provenant d'espèces ou de sous-espèces similaires, peut engendrer des graphes d'assemblage très complexes nécessitant des heuristiques adaptées pour la création et l'élongation des *contigs*. Un autre défi de taille concernant l'assemblage de métagénomes est la quantité massive de données produite par les séquenceurs nécessitant d'être analysées sur des ressources informatiques limitées. Par exemple, le séquençage d'un génome bactérien typique d'une taille de 4 millions de paires de bases (pb) avec une couverture de 50x produira 200 millions de pb, alors qu'un séquençage typique de métagénome à 50 millions de lectures appariées pourra engendrer jusqu'à 10 milliards de pb. Il existe plusieurs assembleurs de métagénomes qui utilisent différentes techniques pour arriver à produire des séquences contiguës malgré la complexité du problème. Tout comme les assembleurs de génomes (sous-section I.3.3), les assembleurs de métagénomes utilisent habituellement l'une des techniques suivantes : le consensus et chevauchement ou les techniques du graphe de De Bruijn [19]. L'un des logiciels les plus populaires pour l'assemblage de métagénomes est MEGAHIT [230]. MEGAHIT utilise les graphes de De Bruijn succinct qui sont une représentation compressée du graphe de De Bruijn original et améliore significativement la mémoire requise pour l'assemblage. Ray Meta est un autre assembleur de métagénomes efficace avec une implémentation axée sur la parallélisation du calcul [34]. En effet, Ray Meta permet de distribuer les différentes étapes d'un assemblage sur plusieurs noeuds de calcul, le rendant adapté aux architectures typiques des superordinateurs. Enfin, metaSPAdes est une couche logiciel par-dessus l'assembleur SPAdes qui ajoute des heuristiques pour résoudre les enjeux typiques aux métagénomes, tels que la divergence de couverture et le mélange d'espèces ou sous-espèces génétiquement similaire [281].

D'autres logiciels d'analyses métagénomiques, non présentés en détail dans cette thèse, incluent MEGAN [175], COGNIZER [37], FMAP [204], ShotMAP [272], et plusieurs autres. L'équipe de recherche ayant développé le logiciel MEGAN sont les mêmes qui ont développé l'aligneur DIAMOND [43], présenté dans la section I.3.2 sur l'identification de séquences par alignement. L'intérêt derrière le développement de DIAMOND était justement pour accélérer

l'identification des courtes lectures d'ADN avec d'imposantes bases de données de protéines pour les annotations de métagénomés.

Le chapitre 2 de cette thèse présente le nouveau logiciel d'identification de protéine kAAmer avec un cas d'utilisation pour la métagénomique. KAAmer présente des résultats supérieurs à DIAMOND en termes d'efficacité pour l'annotation de protéines et est d'ailleurs utilisé dans le chapitre 3 pour l'annotation de métagénomés du microbiote intestinal humain. La prochaine section introduit l'importance du microbiote humain pour la santé humaine et son implication dans différentes conditions pathologiques.

I.5 Microbiote intestinal humain et obésité

I.5.1 Microbiote intestinal humain

Les interactions entre le microbiome et le système immunitaire chez l'humain auraient, de toute évidence, un impact sur son bien-être et pourraient même être associées à plusieurs maladies et conditions pathologiques [250]. Un microbiome peut être défini comme une communauté de microbes caractéristique qui occupe un habitat défini avec des propriétés physico-chimiques distinctes [27]. Il peut même être considéré, à juste titre, comme un organe du corps humain [25]. Non seulement le microbiome inclut les microorganismes qui l'occupent (microbiote), mais aussi leurs activités métaboliques et les éléments génétiques (métagénomiques) qui accomplissent les différentes fonctions souvent essentielles et parfois néfastes à l'hôte ou à l'écosystème dans lequel il habite. La figure I.19 schématise cette définition d'un microbiome et son "théâtre d'activités". L'ensemble des bactéries des microbiotes du corps humain compterait au moins le même nombre de cellules que l'humain possède de cellules humaines [354; 353]. Il va donc sans dire que la compréhension du fonctionnement des microbiotes et de leur impact sur la santé humaine est une question de grand intérêt qui suscite l'engouement des chercheurs, des compagnies pharmaceutiques et de l'industrie agroalimentaire.

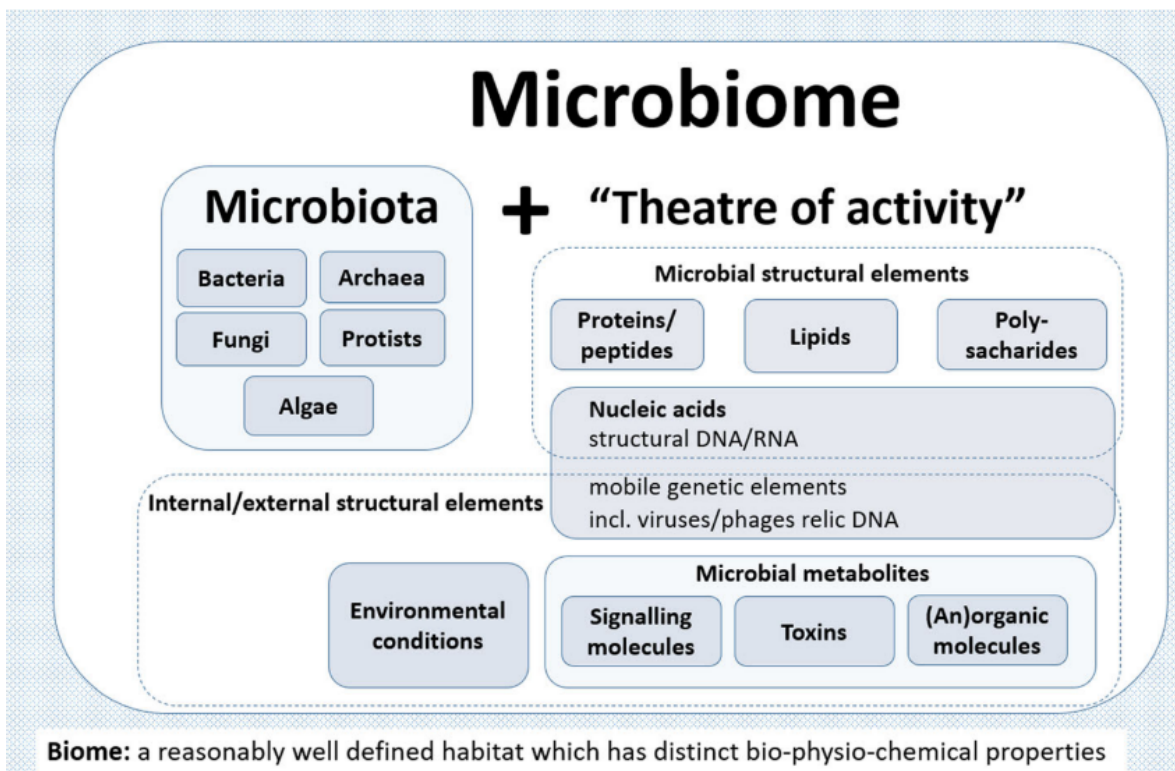


FIGURE I.19 – Illustration des composantes qui définissent un microbiome. Figure issue de Berg et collaborateurs [27]

Plusieurs études épidémiologiques ont analysé les microbiomes du corps humain en relation

avec différentes conditions pathologiques observées chez les individus. La majorité des recherches se sont concentrées sur le microbiote intestinal (voir figure I.20), puisque ses interactions avec l'hôte au niveau du colon suggèrent une importante implication dans différentes activités métaboliques [250]. En effet, les acides gras à chaîne courte (SCFA³⁹), tels le butyrate, le propionate, l'acétate et le pentanoate, lesquels proviennent principalement de la fermentation des fibres alimentaires et de l'amidon résistant [418; 419; 166], seraient une source d'énergie non négligeable pour le métabolisme de l'hôte en plus d'être nécessaire à son homéostasie [131; 330; 331; 32; 47; 77; 97]. Les SCFA seraient aussi, de toute évidence, impliqués dans une signalisation entre l'intestin et le cerveau humain et auraient des effets sur l'appétit et la satiété [389]. Mis à part les SCFA, plusieurs vitamines sont aussi métabolisées par des bactéries du microbiote intestinal [203]. C'est le cas de la vitamine K₂ (ménaquinone) qui réduirait le risque de maladie coronarienne [135] ainsi que des vitamines B5 et B12 qui seraient exclusivement produites par le microbiote et que leur déficience seraient impliquées dans plusieurs conditions comme l'insomnie et certains problèmes neuropsychologiques et hématologiques [14; 142]. Malgré tout, le lien entre la perte des bactéries du microbiote qui produisent ces vitamines et les conditions associées n'a toujours pas été élucidé [203].



FIGURE I.20 – Morphologie de l'intestin humain avec micrographie électronique d'une partie de l'intestin grêle avec des bactéries identifiées en vert. Figure adaptée de Bajzer et al. [22].

Parmi les conditions pathologiques qui ont fait l'objet d'études métagénomiques du microbiote

39. SCFA : *Short-Chain Fatty Acid*

intestinal, on compte notamment les infections à *Clostridium difficile* [263; 11; 99], les maladies inflammatoires chroniques de l'intestin (IBD⁴⁰) [315], la maladie coeliaque (intolérance au gluten) [398], le cancer colorectal [436], la cirrhose du foie [317], le diabète de type 2 [316; 194] et l'obésité [223]. La dysbiose est souvent au coeur de la discussion sur les microbiotes sains et malades et peut être défini comme étant un déséquilibre de l'écosystème des microbes qui constitue le microbiote [335]. Le déséquilibre est souvent caractérisé par un manque de diversité en microbes et de leur contenu génique, une différence importante dans l'abondance des espèces ou encore la présence excessive d'un pathobionte (bactérie pathogène du microbiote), tel que *Clostridium difficile*.

Pasolli et collaborateurs ont réexaminé à l'aide de l'apprentissage automatique les données d'études métagénomiques en relation avec les maladies de cirrhose du foie, du cancer colorectal, de l'inflammation chronique de l'intestin, du diabète de type 2 et de l'obésité [297]. Les résultats de classification, réalisés avec différents algorithmes d'apprentissage, se sont révélés probants principalement pour la cirrhose du foie, la maladie inflammatoire chronique de l'intestin et le cancer colorectal avec des exactitudes atteignant respectivement 87.7%, 80.9% et 80.5%. Cependant, le diabète de type 2 et l'obésité n'ont pas démontré d'aussi bons résultats en termes de classification. La prochaine sous-section I.5.2 introduira l'obésité et l'étude du microbiote intestinal en relation avec cette condition.

I.5.2 Obésité et microbiote

L'obésité est un enjeu épidémiologique très important qui en 2015 touchait globalement plus de 107 millions d'enfants et 603 millions d'adultes [70]. Depuis les années 1975, la prévalence du nombre de personnes obèses a au moins doublé et presque triplé selon les rapports de l'organisation mondiale de la santé (OMS) [422]. Les catégories d'obésités sont définies avec l'indice de masse corporelle (IMC) : surpoids (IMC ≥ 25 kg/m²), obésité 1 (IMC ≥ 30 kg/m²), obésité 2 (IMC ≥ 35 kg/m²) et obésité 3 (IMC ≥ 40 kg/m²) [50]. Il faut cependant noter que l'IMC n'est pas une représentation parfaite de la quantité de graisse corporelle et ne serait pas adapté pour certaines ethnicités tel que pour la population asiatique [141].

L'obésité est aussi un facteur de risque important à plusieurs conditions, telles que le diabète de type 2 [6], les maladies rénales chroniques [324], plusieurs cancers [308] et des troubles musculosquelettiques [291]. Le diabète de type 2 est l'une des conditions les plus dévastatrices reliées à l'obésité [6]. Un lien partagé entre l'obésité et le diabète de type 2 est la résistance à l'insuline [181]. En effet, les tissus adipeux relâcheraient une quantité importante d'acides gras non estérifiés, de glycérol, d'hormones, de cytokines pro-inflammatoires et autres facteurs qui sont impliqués dans la résistance à l'insuline [189]. Somme toute, les deux conditions sont parfois intrinsèquement liées et pourraient être causées par des facteurs croisés comme une

40. IBD : *Inflammatory Bowel Disease*

diète riche en gras trans, une consommation élevée de boissons riche en fructose ainsi que l'inactivité physique [173].

Plusieurs études métagénomiques sur le microbiote intestinal ont été réalisées en vue d'approfondir nos connaissances sur l'obésité et le diabète de type 2. Ce sont d'ailleurs ces deux conditions qui furent les moins conclusives dans les expérimentations d'apprentissage automatique de Pasolli et al. tels que mentionnés précédemment [297]. Le plus préoccupant dans les études métagénomiques en lien avec l'obésité est les résultats contradictoires reportés dans la littérature. En effet, l'une des mesures souvent rapportées dans les analyses de métagénomes obèses est le ratio des deux phyla les plus abondants du microbiote intestinal (*Bacteroidetes* et *Firmicutes* [315]), une donnée qui ne concorde pas entre toutes les études [223]. L'étude de Ley et collaborateurs, de 2005, a observé à l'aide du séquençage du gène de l'ARN 16S chez 19 souris une proportion plus élevée en *Firmicutes* et moins élevée en *Bacteroidetes* chez les souris obèses comparativement à celles sans la condition [228]. Ensuite en 2006, Ley et collaborateurs reproduisirent une expérimentation similaire, mais cette fois chez 12 sujets humains atteints d'obésité combinés à 12 sujets témoins [229]. Tout comme leur étude chez les souris, ils reportèrent un ratio en faveur des *Firmicutes* chez les sujets obèses lorsque comparés au groupe contrôle. Une autre étude de Turnbaugh et al. en 2006, explorait aussi les métagénomes du microbiote intestinal des souris, mais cette fois à l'aide du séquençage par *shotgun* et reportait aussi un ratio en faveur des *Firmicutes* dans les microbiotes des sujets obèses [394]. Toutefois, en 2008, Duncan et collaborateurs, dans leur étude métagénomique basée sur l'ARN 16S sur 23 sujets humains obèses 14 sujets témoins, ne trouvèrent aucune évidence de corrélation entre les deux phyla et la condition d'obésité [100]. Finalement, l'étude métagénomique de Schwartz et al. de 2010 était aussi basée sur l'ARN 16S et elle comprenait 98 individus, dont 30 avec un IMC normal, 35 avec un IMC de surpoids et 33 avec un IMC lié à l'obésité. Contrairement à ce qui avait été rapporté dans les études antérieures, leurs résultats par rapport au ratio de *Firmicutes* et *Bacteroidetes* présentèrent un enrichissement en *Bacteroidetes* et conséquemment une diminution en abondance relative des *Firmicutes* chez les individus obèses et en surpoids [349]. Par ailleurs, ils reportèrent un changement qui serait autant sinon plus significatif que la composition en bactéries, soit la quantité de SCFA produite par le microbiote intestinal. Ils ont en effet constaté une concentration de SCFA beaucoup plus importante dans les matières fécales des individus obèses. Les causes de cette augmentation en SCFA pourraient être variées, avec comme explication possible une absorption diminuée des SCFA par les cellules épithéliales du colon, mais sembleraient quand même corrélées à un certain niveau avec la composition bactérienne des métagénomes.

Pour explorer davantage la relation entre la composition taxonomique et fonctionnelle des métagénomes du microbiote intestinal et de la condition d'obésité, nous avons réalisé une étude sur le sujet avec les données de séquences de métagénomique par *shotgun* de 640 individus présentant des IMC normaux, de surpoids et d'obésités. Le chapitre 3 présente donc cette

étude métagénomique qui confirmera les résultats de Schwartz et al. en ce qui a trait le ratio de *Firmicutes* et *Bacteroidetes* chez les obèses, en plus de présenter une analyse fonctionnelle des différentes fonctions encodées par les métagénomes ayant potentiellement un lien avec la condition d'obésité.

I.6 Hypothèses et approches méthodologiques

La complexité et la taille grandissante des études génomiques et métagénomiques en lien avec l'analyse de génomes bactériens nécessitent sans cesse le perfectionnement de logiciels adaptés pour réaliser les analyses bio-informatiques. En effet, historiquement le séquençage ne visait habituellement qu'un gène ou de petite portion de génomes. Aujourd'hui, le séquençage de génomes complets, d'une multitude de génomes d'une même espèce ou encore de plusieurs métagénomes avec le séquençage par *shotgun* sont couramment utilisés. Les avancées techniques dans les méthodes de séquençage et la diminution drastique de son coût ont donc beaucoup apporté au domaine de la génomique. L'évolution des méthodes en génomique requiert aussi une évolution des logiciels bio-informatiques pour s'adapter à la complexité des analyses. Avec l'adoption universelle du séquençage et le dépôt souvent obligatoire des séquences génomique dans les bases de données publiques (NCBI, EBI, DDBJ) pour fin de publication dans les journaux scientifiques, les ressources disponibles pour les chercheurs en génomique sont incommensurables. Plusieurs logiciels et bases de données en bio-informatique ont donc vu le jour au cours des années pour améliorer les performances et la qualité des analyses en génomique tout en faisant face à la quantité massive de données. Durant ce projet de doctorat, nous avons donc exploré plus en profondeur les techniques d'identification de séquences génomique à l'aide des k -mers et de leur application dans des études métagénomiques en relation avec une condition néfaste à la santé humaine.

La première hypothèse de recherche consistait à vérifier s'il était possible d'obtenir une comparaison de génomes, depuis leur simple contenu en k -mers de séquences d'ADN, avec des résultats analogues aux comparaisons génomiques standards comme le pourcentage moyen d'identités ou la phylogénie, mais sans nécessiter d'alignements de séquences. De plus, la comparaison génomique pourrait permettre d'appliquer des filtres avec des bases de données sur des gènes spécialisés pour estimer la disposition génotypique des bactéries à des phénotypes d'intérêt clinique. Le chapitre 1 présente les travaux de recherches liés à cette problématique ayant pris forme avec l'implémentation du logiciel Ray Surveyor.

La deuxième hypothèse était de tester s'il était possible de développer un logiciel pour l'identification de séquences protéiques, basé sur des k -mers d'acides aminés, qui serait plus performant que les logiciels existants pour l'identification de protéines avec un haut degré d'homologie. La méthode prendrait la forme d'un engin de base de données clé-valeur qui offrirait une flexibilité sur l'inclusion des annotations de protéine. Elle serait aussi optimisée pour les nouvelles générations de stockage rapide (SSD⁴¹) et pourrait être hébergée dans l'infonuagique de manière permanente et interrogée à distance via un API. Le chapitre 2 présente les travaux sur kAAmer, un logiciel qui implémente les fonctionnalités nécessaire à l'hypothèse de recherche.

Enfin, la dernière partie de cette thèse avait pour but d'appliquer les méthodes développées

41. SSD : *Solid-State Drive*

durant le doctorat dans l'analyse de métagénomies. Formellement, la troisième hypothèse de recherche visait à valider si les deux logiciels Ray Surveyor et kAAMer seraient en mesure de produire des résultats viables dans une analyse métagénomique du microbiote intestinal humain en lien avec la condition d'obésité. Plusieurs études sur le sujet avaient d'ailleurs reporté des résultats, parfois contradictoires, en ce qui a trait les abondances relatives des phyla *Bacteroidetes* et *Firmicutes*. Les travaux de recherches incluaient donc aussi la validation des différents résultats reportés antérieurement avec une cohorte de métagénomies du microbiote provenant de plusieurs études distinctes. Le chapitre 3 présente ces travaux d'analyses métagénomiques sur l'obésité à l'aide des méthodes implémentées durant ce doctorat.

Chapitre 1

Comparaison phénétique de génomes procaryotes basée sur les k-mers

Titre original

Phenetic Comparison of Prokaryotic Genomes Using k-mers [90].

Journal

Molecular Biology and Evolution, Volume 34, Issue 10, October 2017, Pages 2716–2729, <https://doi.org/10.1093/molbev/msx200>

Auteurs

Maxime Déraspe, Frédéric Raymond, Sébastien Boisvert, Alexander Culley, Paul H. Roy, François Laviolette, Jacques Corbeil

MD et FR sont co-premiers auteurs. FR et JC sont coauteurs séniors.

1.1 Résumé

Les études génomiques sont plus en plus complexes et nécessitent de nouvelles techniques d'analyse capable de gérer la quantité massive de données produites par le séquençage à haut débit. Nous avons développé un nouveau logiciel nommé Ray Surveyor pour permettre la comparaison de génomes basée sur leur contenu en k-mers ainsi que la reconstruction d'arbres phénétiques. Nous avons validé la méthode sur des données simulées ainsi que sur des données réelles provenant d'études déjà publiées portant sur les espèces *Streptococcus pneumoniae* et *Pseudomonas aeruginosa*. Nous avons aussi examiné la relation entre plusieurs éléments génétiques et la structure de population de génomes bactériens. En effet, en comparant les groupements de génomes complets et ceux de génomes filtrés sur des catégories de gènes spécifiques, on peut déterminer une corrélation entre la structure des groupements. Cette corrélation permet d'évaluer l'implication d'une catégorie de gènes sur une population de génomes d'une même espèce. Nous avons appliqué cette méthode sur 42 espèces de bactéries pour déterminer l'importance de cinq caractéristiques génomiques importantes chez les bactéries. Par exemple, la résistance intrinsèque aux antibiotiques chez *P. aeruginosa* est démontrée par les résultats probants de corrélation des gènes associés à la résistance aux antibiotiques et la structure entière de sa population génomique. Une vue étendue du pangénome des bactéries illustre aussi différents niveaux d'interaction entre la structure de leur population et les catégories de gènes associés à la résistance aux antibiotiques, aux bactériophages, aux plasmides et aux éléments mobiles.

1.2 Abstract

Bacterial genomics studies are getting more extensive and complex, requiring new ways to envision analyses. Using the Ray Surveyor software, we demonstrate that comparison of genomes based on their k-mer content allows reconstruction of phenetic trees without the need of prior data curation, such as core genome alignment of a species. We validated the methodology using simulated genomes and previously published phylogenomic studies of *Streptococcus pneumoniae* and *Pseudomonas aeruginosa*. We also investigated the relationship of specific genetic determinants with bacterial population structures. By comparing clusters from the complete genomic content of a genome population with clusters from specific functional categories of genes, we can determine how the population structures are correlated. Indeed, the strain clustering based on a subset of k-mers allows determination of its similarity with the whole genome clusters. We also applied this methodology on 42 species of bacteria to determine the correlational significance of five important bacterial genomic characteristics. For example, intrinsic resistance is more important in *P. aeruginosa* than in *S. pneumoniae*, and the former has increased correlation of its population structure with antibiotic resistance genes. The global view of the pangenome of bacteria also demonstrated the taxa-dependent interaction of population structure with antibiotic resistance, bacteriophage, plasmid, and mobile element k-mer data sets.

1.3 Introduction

Genomic data sets are continuously increasing in size and a single study now contains hundreds to thousands of samples that must be rigorously compared and clustered [271; 405]. Large-scale genomic projects such as the 1,000 genomes project [362], the Human Microbiome Project [176] or any recent epidemiological studies of outbreaks [105; 365; 138] rely on comparative genomics and large scale phylogenies to uncover underlying biological patterns and trends. Nowadays, sequenced genomes are compared based on conserved genes, polymorphic positions and/or annotations (16S rRNA, *rpoB*, *atpB*, etc.) [298]. For example, multilocus sequence analysis (MLSA) uses the sequences of housekeeping genes to construct phylogenies [139]. On a larger scale, phylogenomics often compare genomes using the conserved genes of the population under study [302]. Another common approach for whole genome comparison is the Average Nucleotide Identity (ANI) that relies on sequence alignments in order to determine the percentage of similarity between genomes [208]. Researchers are thus often interpreting their results solely based on a comparison of the shared features of their samples, an approach that may omit important genomic determinants that could better characterize and discriminate subpopulations or phenotypes [391]. Indeed, the accessory or dispensable genome can be responsible for important phenotypes such as antibiotic resistance, adaptation to specific environments or colonization of different hosts [260]. Genes acquired by horizontal gene transfer (HGT) are not measured by traditional methods that use conserved genes to compute evolutionary distance between bacteria. Given the importance of the accessory genome in pathogen traits, such as virulence and antibiotic resistance, it is of interest to have analytical tools capable of comparing thousands of genome sequences without reducing analysis to conserved features.

K-mer-based methodologies are not new and have attracted researchers' interest for quite a while now [401; 367; 159]. It is the gold standard for short read assemblies with De Bruijn graphs [72; 34] and there are several highly efficient k-mer counters, like MSPKmerCounter [235], DSK [326], and KMC2 [83]. Alignment-free sequence comparisons have been studied in numerous ways and are competitive with alignment-based methods in terms of accuracy while being generally computationally more efficient [251; 132; 289]. They have also been used for the comparison of assembled microbiomes [321] and proved to be an important tool in the phylogenetic analysis toolbox [314; 413]. Comparison of k-mer content can also be combined with machine learning algorithms to predict phenotypes such as antibiotic resistance [98].

In this work, we evaluated whether k-mers can be used to rapidly and accurately compare large collections of genomes. With this approach, genomes are clustered based on the similarity of their complete sequence by counting the total number of shared k-mers, including the accessory genome. In addition, we tested the hypothesis that it is possible to characterize populations of genomes based on specific features using presence/absence of k-mers related to these features. To do so, we filtered genome sequences by selecting only k-mers that were also present in a

reference sequence data set, and then compared the clustering of whole genomes against the filtered genomes. The purpose of the filtered data set is to establish a functional set of genes with common characteristics. We then suggest that if genome clustering based on specific gene functions restores the population structure based on whole genomes, this functional set of genes is linked to the structure of the population under study. This suggests that the functional set of genes could have a conserved function in the population and presumably a selective pressure similar to the whole genomes, for example. On the basis of this logic, we explored this relationship by comparing a large number of bacterial genomes with several gene sequence data sets, each one representing a different functional gene category. We used reference sequence data sets of antibiotic resistance genes (ARG), insertion sequences, plasmids, bacteriophages and biosynthetic gene clusters (BGC) and observed their relationship with genome population structure for different bacterial species. This approach is implemented in the Ray Surveyor software, which is built on top of the scalable Ray framework [33; 34]. The defining feature of Ray Surveyor is the ability to compare whole genomes based on their complete set of k-mers along subsets of their k-mers, filtered with other sequence data sets. Ray Surveyor allowed us to determine how the five genetic element categories tested are linked with the population structure of 42 species of bacteria.

1.4 Results and Discussion

Validation with Simulated Genome Populations

To overcome possible uncertainties introduced by real genome data sets, we started by generating random phylogenetic trees [213; 150; 31] and simulating genome sequences from these trees [370]. Three different branch lengths were used to simulate tree structures in order to measure the impact of this parameter on the clustering methods used in Ray Surveyor analyses. The branch lengths were computed using an exponential distribution, which yielded an average depth of $\log_2(n)$ with n being the number of genomes in the tree, 100 in our case. For each average branch length, ten random trees were computed to evaluate reproducibility. Sequences of one million nucleotides were produced for each simulated genome in the phylogenies in the form of an alignment, using Pyvolve [370].

The three branch lengths we examined were chosen to model bacterial populations of within-species genomes (0.001), within-genera genomes (0.005), and interspecies genomes (0.01). This assumption was based on the ANI of all simulated trees. The ANI cutoff to distinguish bacterial species is estimated to be between 93 and 96% ANI [334]. Consequently, trees with an average branch length of 0.001 (average ANI = 98.3%) are akin to intraspecies data sets and branch lengths of 0.01 (average ANI = 85.4%) to interspecies data sets. An average branch length of 0.005 corresponds to an ANI of 92.1% between all pairs of genomes, with 56.5% of them being below 93%. Therefore, in trees with an average branch length of 0.005, half of the genomes belong to the same bacterial species whereas the other half belongs to different species from the same genera. Although these cut-offs do not apply to all bacterial species, they generally reflect the current state of the NCBI taxonomy and they allow the evaluation of the influence of strain diversity on comparative genomics methods.

To allow comparison of Ray Surveyor clusters with phylogenies, we took the distance matrices derived from the simulated trees and generated a dendrogram by hierarchical clustering with the UPGMA linkage method. Similarly, the k-mer Gram matrices generated with Ray Surveyor were transformed into distance matrices upon which hierarchical clustering dendrograms were computed. Those dendrograms are referred to as phenetic trees throughout the manuscript. The cophenetic correlation coefficient (CCC) [366] was then used to assess how Ray Surveyor phenetic trees correlated with the simulated phenetic trees. The CCC in our case measures how well two phenetic trees preserve the pairwise distances between all pairs of genomes. We tested the impact of four distance metrics on the transformation of the Ray Surveyor Gram matrix using Euclidean, cosine, correlation and Canberra distances. Ray Surveyor analyses were also performed with k-mer lengths ranging from 11 to 101 nucleotides to evaluate their impact on accuracy.

The cophenetic correlation results from Ray Surveyor analyses were affected by the average pairwise phylogenetic distance of genomes and the k-mer lengths used in the analysis (fig.

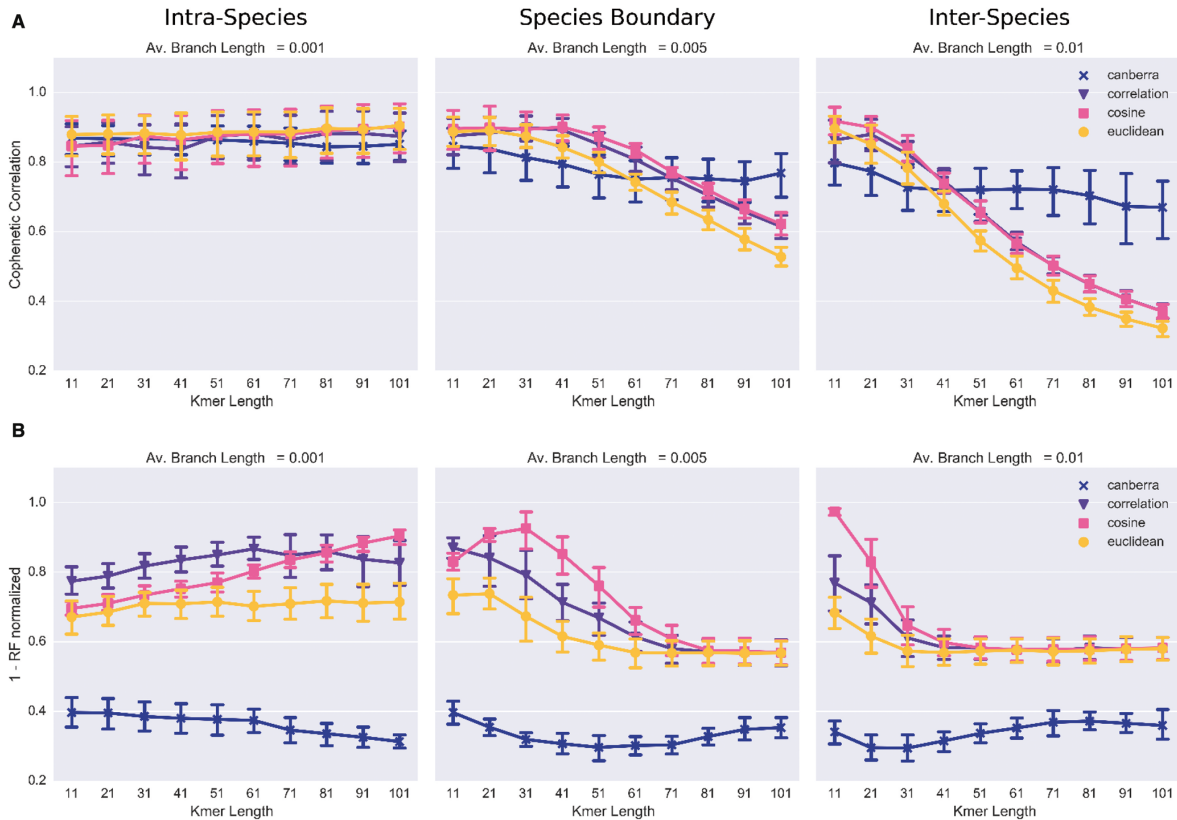


FIGURE 1.1 – Evaluation of simulated genome populations with Ray Surveyor. Colors and symbols represent the distance metrics used to transform the Ray Surveyor’s Gram matrix into a distance matrix. Each column represents a different evolutionary distance between the genomes, based on the average branch length and bacterial species definition. Ten replicates were performed for each point. First row (A) is the cophenetic correlation between the reference phylogeny and the phenetic tree. Second row (B) is the Robinson–Foulds metric between the reference phylogeny and the Ray Surveyor derived tree.

1.1A). Indeed, CCCs were higher for intraspecies genome populations (lower average pairwise distance) and were only slightly affected by k-mer length or distance metrics. When genome populations grew more distant, crossing the species boundary, CCCs decreased with increasing k-mer length. By comparing distance metrics used to construct phenetic trees based on Ray Surveyor results, we observed that Euclidean, cosine, and correlation distances behaved similarly on simulated genome populations (see Materials and Methods). The Canberra distance provided lower CCC for more closely related genomes, but it was less affected by more heterogeneous populations when the k-mer lengths were increased. This result is likely due to the fact that the Canberra distance is more tolerant of low absolute values (the number of shared k-mers), as observed by Loureiro et al. (2004) [241]. For a control, we produced alignment-based phylogenies of the simulated sequences that had average CCCs of 0.98 for branch lengths of 0.001, 0.97 for 0.005 and 0.99 for 0.01.

In order to test the capability of Ray Surveyor to restore good topologies for phylogenetic trees,

we also computed a Neighbor-Joining tree for all the distance matrices. In this comparison, we used the original simulated phylogenetic tree against those derived from Ray Surveyor. The Robinson–Foulds (RF) metric allows a comparison of unrooted phylogenetic trees, essentially by measuring the number of changes required to align two trees together by transforming one tree into the other [328]. Similar to the cophenetic correlation, the RF results varied with sequence diversity and k-mer length (fig. 1.1B). For the intraspecies genome populations (branch length = 0.001, average ANI = 98.3%) longer k-mer length performed better and peaked with the 101-mers and the cosine metrics. At the species boundary, the cosine distance metrics yielded the best topological trees with 31-mers. When comparing genomes of different species (branch length = 0.01, average ANI = 85.4%), a k-mer length $<$ of 31 yielded better topological trees for the cosine, correlation and Euclidean metrics.

On the basis of these results and on the literature, the choice of k-mer length can be seen as a trade-off between sensitivity and specificity [289]. Evolutionarily distant genomes require shorter k-mers to get a good signal (sensitivity) whereas more similar genomes benefit from larger k-mer lengths for more specificity. Moreover, previous studies have shown the efficiency of 31-mers in genome clustering [261] and the robustness in bacterial metagenome profiling [34] when this length of k-mer is used. For the following analyses on real genome data sets, we selected a length of 31-mers, which offers a compromise between sensitivity and specificity for both intraspecies and interspecies comparison. We also focused our analyses on the cophenetic correlation for the phenetic trees, since we needed to characterize genomes based on specific genetic elements rather than finding their ancestral history.

Population Scale Genomics with k-mers

This section aims to benchmark the application of the Ray Surveyor genome comparison in comparative genomics projects and to assess how it performs on microbial populations of different scales. As a first step, we validated that k-mer-based phenetic trees accurately reflected previously determined phylogenies based on publicly available comparative genomic studies of *Streptococcus pneumoniae* and *Pseudomonas aeruginosa* (fig. 1.2). For *P. aeruginosa*, 387 genomes were taken from a study by Kos et al. [211] (1.2A). For *S. pneumoniae*, a first data set of 616 genomes from Croucher and collaborators was used, along with a second data set comprising 173 genomes previously studied by Hilty and collaborators to investigate the difference between encapsulated and nonencapsulated pneumococci [75; 96; 167]. Whole genome phylogenies were obtained from the authors for the Kos and the Croucher data sets, while the phylogeny for the Hilty collection was built using 602 conserved genes. We calculated the cophenetic correlation between the phenetic trees (hierarchical cluster dendrograms) created using Ray Surveyor and the derived phenetic trees from the phylogenies for these three data sets (fig. 1.2A). All four distance metrics (see Materials and Methods) performed above 0.91 CCC on *P. aeruginosa*, with the Canberra distance yielding the highest CCC of 0.97. Corre-

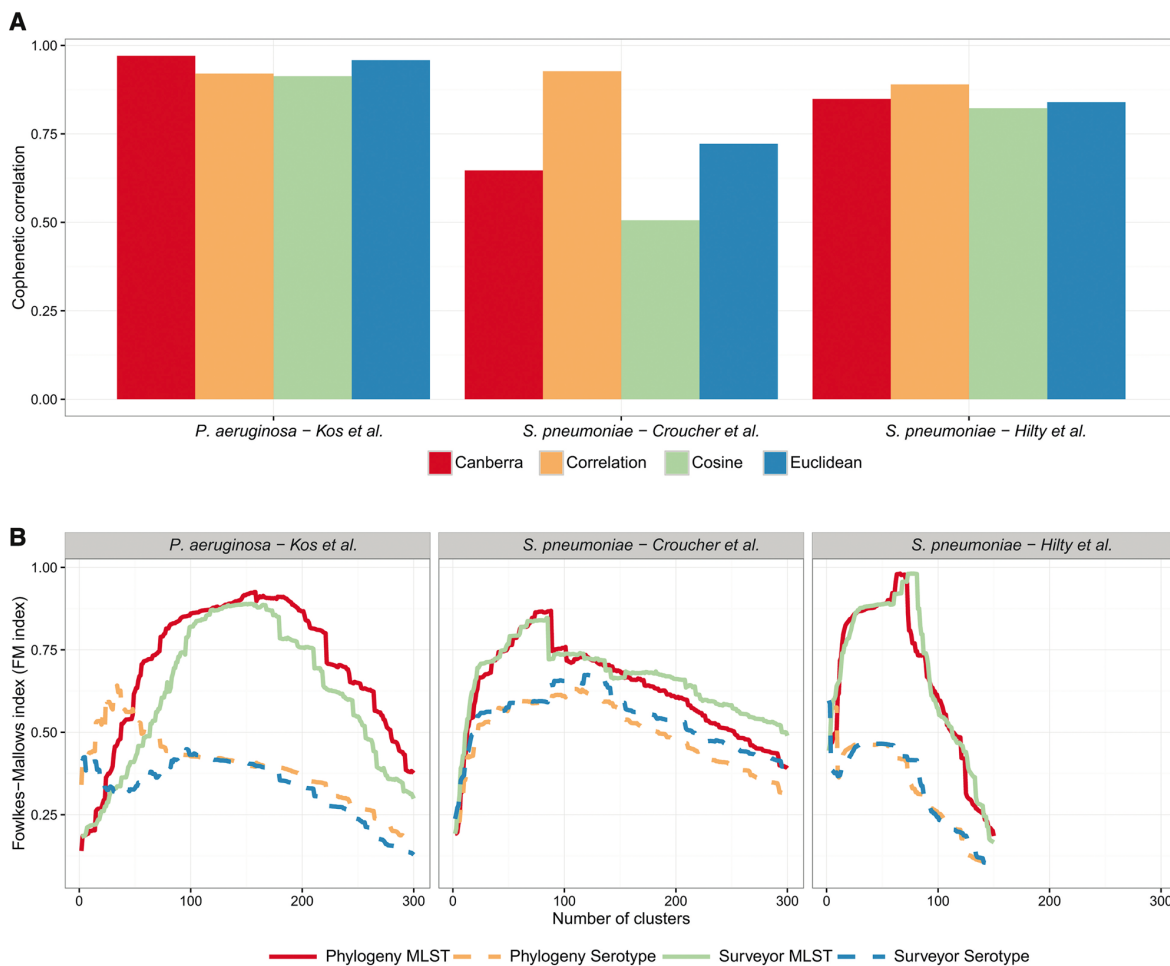


FIGURE 1.2 – Comparison of phenetic trees created using Ray Surveyor to phylogenies calculated using conserved genomes or marker genes for *Pseudomonas aeruginosa* and *Streptococcus pneumoniae*. (A) Cophenetic correlation between alignment-based phylogeny and phenetic trees calculated using four different distance metrics. (B) Fowlkes–Marlows index comparing clustering done using Ray Surveyor (correlation distance metric) and phylogeny compared with classification based on multiple locus sequence typing or serotypes.

lation distance had the highest CCC (0.92) compared with other distance metrics (<0.75) for *S. pneumoniae*. The Hilty and collaborators data set of *S. pneumoniae* genomes was tested and provided 0.89 CCC between correlation distance based on k-mers and the core genome phenetic tree. Heatmaps representing the clustering based on the distance between isolates of the Croucher and Kos data sets are shown in supplementary figures 1 and 2, Supplementary Material online.

This approach can also be used to quickly add a new genome to an existing phylogeny. For example, we added the recently sequenced genome of *P. aeruginosa* strain E6130952 to the Kos et al. genome collection (CP020603.1 [<https://www.ncbi.nlm.nih.gov/nucleotide/CP020603>]; last accessed July 19, 2017]; supplementary fig. 3, Supplementary Material online). This pa-

thogenic strain was isolated from a patient with respiratory failure and was resistant to all tested antibiotics, including colistin [426]. The closest isolate in the phylogeny (AZPAE14730) was also resistant to levofloxacin, meropenem, and amikacin, but not to colistin [211]. Both strains have a similar genome size and share 97% of their k-mers.

In epidemiological studies, genomes are often classified based on experimentally derived categories such as multilocus sequence typing or serotypes. The Fowlkes–Mallows index (FMI) allows calculation of the similarity between two clusterings [123] and can be used to compare clustering based on k-mers or phylogeny to categorical information of clinical relevance. Thus, we used this metric to quantify the concordance between the clusters generated with Ray Surveyor or with phylogeny to metadata associated with genomes. Therefore, we calculated the FMI between clustering based on the phylogenetic and phenetic trees of *P. aeruginosa* and *S. pneumoniae* when compared with MLST and serotype genome classification, for a range of 2 to N clusters (fig. 1.2B). Phylogenetic genome comparison and k-mer-based comparison provided similar results when compared with MLST or serotype categorization. The highest divergence in FMI between phylogeny and k-mers was <5%. Similarity with MLST was higher ($\geq 85\%$) than similarity with serotype ($\leq 67\%$), suggesting that MLST is more related to complete genome phylogeny than serotype. Indeed, in *S. pneumoniae*, the capsular operon can be modified through capsular switching, a process that decouples serotypes from the core and accessory genomes [12]. In the Hilty data set, genomes from different strain types could be associated within the category of nonencapsulated *S. pneumoniae*, thus explaining the low FMI of serotypes in comparison to the near-perfect FMI obtained when benchmarking against MLST results.

In order to explore Ray Surveyor’s capacity to work with a large number of distantly related genomes, we created a data set of 2,429 complete genomes from 30 phyla in the domain Bacteria. The 2,429 bacterial genomes from which this data set was derived were selected in order to limit the bias caused by a relative overrepresentation of certain genomes in the public database, such as laboratory strains of *Escherichia coli* or clonal isolates from epidemiological studies. We compared the phenetic tree built with these genomes using Ray Surveyor to the 16S rRNA phylogenetic tree of these strains. Canberra distance was the best performing metric (0.69 CCC compared with <0.10 for other distance metrics) for the tree of 2,429 bacterial genomes, most certainly because of the low number of shared k-mers between distant genomes (fig. 1.3A). We also used the FMI to compare 16S phylogenetic and Ray Surveyor phenetic trees to the taxonomical classification of genomes at the family rank based on the NCBI taxonomy. Although the NCBI taxonomy may not always be in line with other taxonomies, it provides a convenient way to perform taxonomy-related analyses with genomic sequences obtained from NCBI [111; 23]. When comparing the classification of 2,429 genomes from 262 bacterial families to genome-based clustering, the peak FMI was 67% for k-mers (469 clusters) compared with 68% for 16S phylogeny (310 clusters; fig. 1.3B). While these methods had similar correlations

with current NCBI taxonomy at the family rank, we also observed that the accuracy of clusters was influenced by the number of genomes within each bacterial family (fig. 1.3B). When considering only bacterial families represented by at least 20 genome sequences (39 families), k-mers had a maximal FMI value of 78% (at 167 clusters) compared with phylogeny which had a maximal value of 77% at 107 clusters. In contrast, when considering families represented by <20 genomes (223 families), FMI was 62% for k-mer analysis (378 clusters) compared with 71% for 16S rRNA phylogenetic trees (451 clusters). The discrepancies between 16S rRNA phylogeny and k-mer-based clustering were mainly associated with regions where only a small number of genomes were included in the analysis. Additionally, the low count of shared k-mers between these small groups of genomes and the rest of the taxa makes it hard to find common ancestors and thus infer their correct placement in the final dendrogram. Hence, efficient clustering of phylogenetically distant bacteria that share a nonsignificant amount of k-mers would require more intermediate genomes to effectively drive the hierarchical clustering and a shorter k-mer length to get more signal.

To investigate the relationship between traits and genome clustering, quantitative and qualitative metadata can be plotted against a phenetic tree. For example, supplementary figure 4, Supplementary Material online, plots a phenetic tree of 2,429 bacterial genomes versus their GC-content and their taxonomic class rank. In this representation, differences in GC-content seem related to the taxonomical classification. Because phenetic trees do not rely on sequence alignments, we cannot correct for GC-content or codon bias using substitution models or other methods, as suggested in the literature [267]. Therefore, we do not expect branch lengths, generated using our k-mer approach, to be representative of evolutionary distance. The clustering of high taxonomic rank could also be biased by GC content [267]. At the k-mer level, differences in GC-content and codon usage should negatively affect k-mer similarity. Indeed, k-mer similarity is expected to decrease quickly as the number of mismatches increase. Previous studies have shown that the type of environment and particular lifestyles of the bacteria is related to genomic GC-content and codon usage [121; 38; 222]. Differences in ecological niches are also reflected in the accessory genome, which can lead to large differences in k-mer content [260].

Comparing Genomes Based on Specific Traits

Not all genes within a genome have the same association with the evolutionary story of a species as inferred from phylogeny [216]. For example, genes acquired by HGT may not be linked to the phylogeny of a species and may have been acquired independently by different strains, for example, genes from mobile elements, bacteriophages or plasmids [307]. Resistance genes as well as secondary metabolite operons [95] can also be disseminated by HGT [295].

In order to investigate HGT patterns in our data set, we developed an approach to quantify how the phenetic tree generated using a subset of k-mers reflected the tree generated using

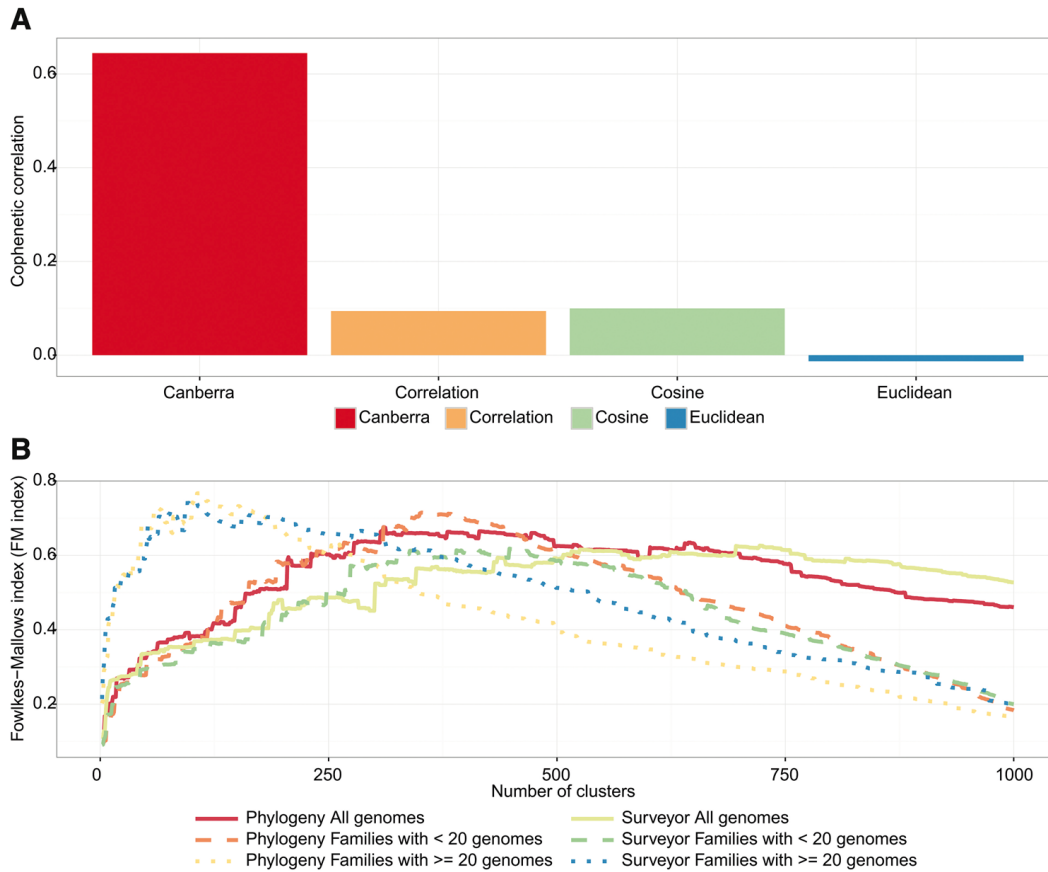


FIGURE 1.3 – Comparison of phenetic trees created using Ray Surveyor to phylogeny based on 16S gene sequence for 2,429 bacterial genomes. (A) Cophenetic correlation between alignment-based phylogeny and phenetic trees calculated using four different distance metrics. (B) Fowlkes–Marlows index comparing clustering done using Ray Surveyor (correlation distance metric) and phylogeny compared with taxonomical classification at the family rank.

the total k-mer content of a genome. We hypothesize that if the two trees are correlated, the group of k-mers is linked to the phylogeny of the studied population. Conversely, the absence of correlation indicates independence between the whole genome population and the filtered genome population. The first steps to conduct the analyses are similar to the ones' explained in the two previous sections. We first calculated a Gram matrix of shared k-mers for all pairs of genomes. For each population two Gram matrices were produced, one with the total count of shared k-mers between the genomes and the second containing only the count of shared k-mers included in the filtering data sets. We then generated a distance matrix for the complete and filtered Gram matrices using the Canberra distance, which we chose in order to reduce bias caused by samples with a limited number of filtered k-mers. Phenetic trees were then built using UPGMA clustering on the distance matrices. We aligned the heatmaps of the clusters based on the whole genome phenetic tree to visualize its similarity with the filtered phenetic tree. In addition, the correlation between phenetic trees based on complete k-mer content and filtered k-mer sets was quantified using CCC. A coefficient of 0 indicates the absence of

correlation whereas a coefficient of 1 indicates perfect cophenetic correlation between selected k-mers and complete genomes, thereby suggesting that these k-mers are associated with the phylogeny of the population.

In our initial analysis, we further investigated genome populations of *S. pneumoniae* and *P. aeruginosa* and the 2,429 bacterial genomes using subsets of k-mers that could be acquired through HGT and may have an impact on the evolution of bacterial species. We used five filtering data sets : mobile elements (insertion sequences), resistance genes, bacteriophages, plasmids, and BGC. The filtering analyses were produced using the strict inclusion of k-mers from the filtering data sets. However, for the plasmids filtering, we also excluded the k-mers from the resistance genes and mobile elements data sets as these genetic elements often co-appear on plasmids and chromosomes. As represented in figure 1.4, the coherence between the heatmaps based on filtering and those based on complete genomes, also expressed quantitatively by the cophenetic correlation, is different between filtering data sets and genome collections. *Streptococcus pneumoniae* showed low (0.28 CCC) correlation between antibiotic resistance k-mers and complete genome clustering. BGC (0.48 CCC) and plasmids (0.52 CCC) had moderate correlation with complete genome clustering. The genomes harbor on average 403 and 5,636 k-mers for BGC and plasmids, respectively, suggesting sequences from these origins are not widely abundant in the species, although they are correlated with the structure of the population. *Streptococcus pneumoniae* does not frequently harbor plasmids, which is reflected in the count of k-mers related to these genetic elements [332]. The lack of characterized BGC from the species in the filtering data set could also have an impact on the moderate correlation. In contrast, the *P. aeruginosa* phenetic trees based on resistance genes (0.98 CCC) and BGC (0.92 CCC) were highly correlated with phenetic trees based on the whole genome. The number of shared k-mers associated with the two filtering data sets was on average 41,240 and 63,820 k-mers, respectively. Similar results were obtained on 71 genomes of *P. aeruginosa* downloaded from the PATRIC database [409], which included some environmental samples, and on 500 *P. aeruginosa* genomes randomly selected from NCBI (supplementary fig. 5, Supplementary Material online). In the case of the 2,429 bacterial genomes data set, the whole genome phylogeny was highly correlated with plasmids and BGC. The overall relationship between representative taxa in the domain Bacteria was not distinctively defined by resistance genes, which are broadly distributed in the microbial tree of life and can be associated with HGT [262].

In order to dissect the relationship between bacterial pathogens and the five filtering data sets, we applied the methodology described above to 42 bacterial species for which at least 100 genomes were available in the NCBI RefSeq database (fig. 1.5). These taxa are associated with human infections, with the exception of *Lactobacillus plantarum* which is found in fermented food [399]. Our hypothesis is that high cophenetic correlation of clustering between complete and filtered k-mer content is a good indicator of how the tested elements are related to the

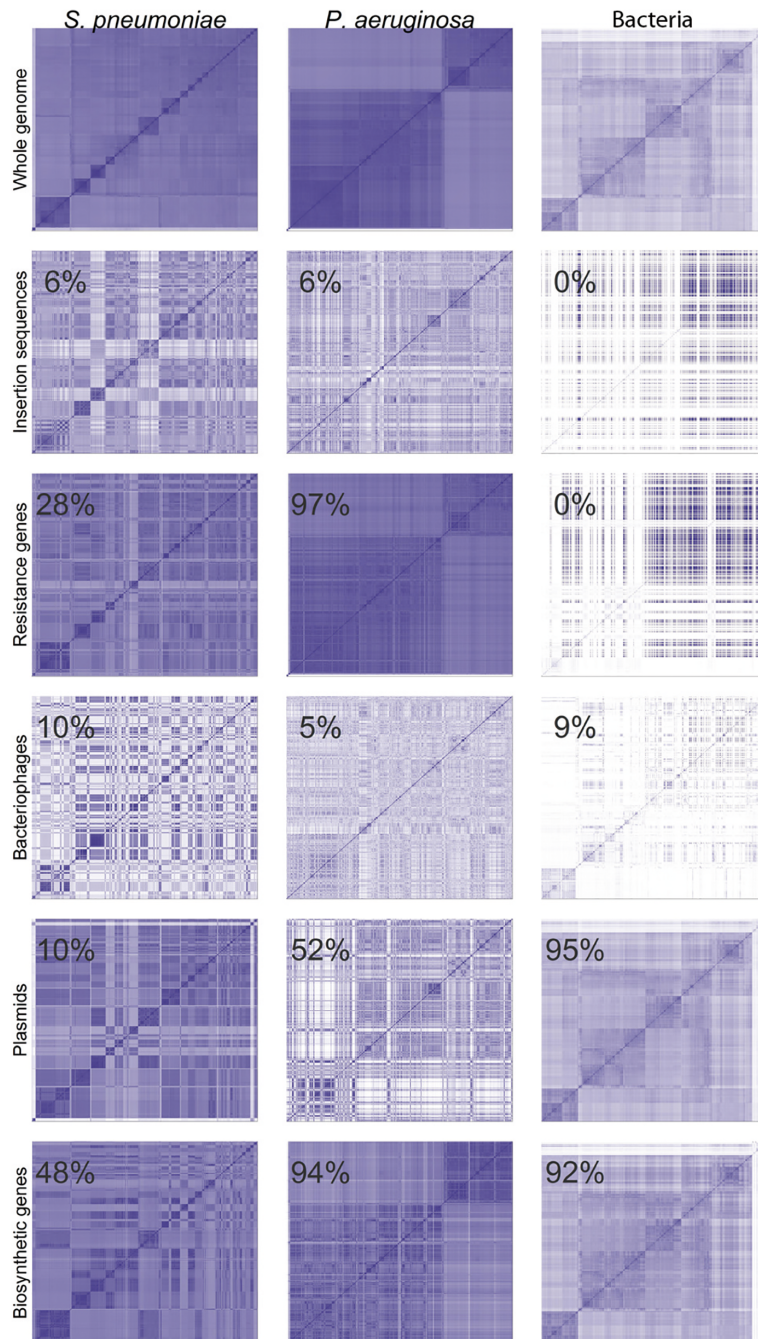


FIGURE 1.4 – Comparison of the relationship between strains when genome sequences are filtered using one of five filtering data sets for *Streptococcus pneumoniae*, *Pseudomonas aeruginosa* and the 2,429 representative bacterial genomes. The Heatmap represents the Canberra distance between genomes collated on a subset of k-mers. The X and Y axis of the heatmap are genomes ordered based on hierarchical clustering of the complete genome. The number in top left corner of heatmaps is the cophenetic distance, expressed in percentages, between filtered data sets and whole genome phenetic tree. The darker the shade of blue, the higher the similarity between samples.

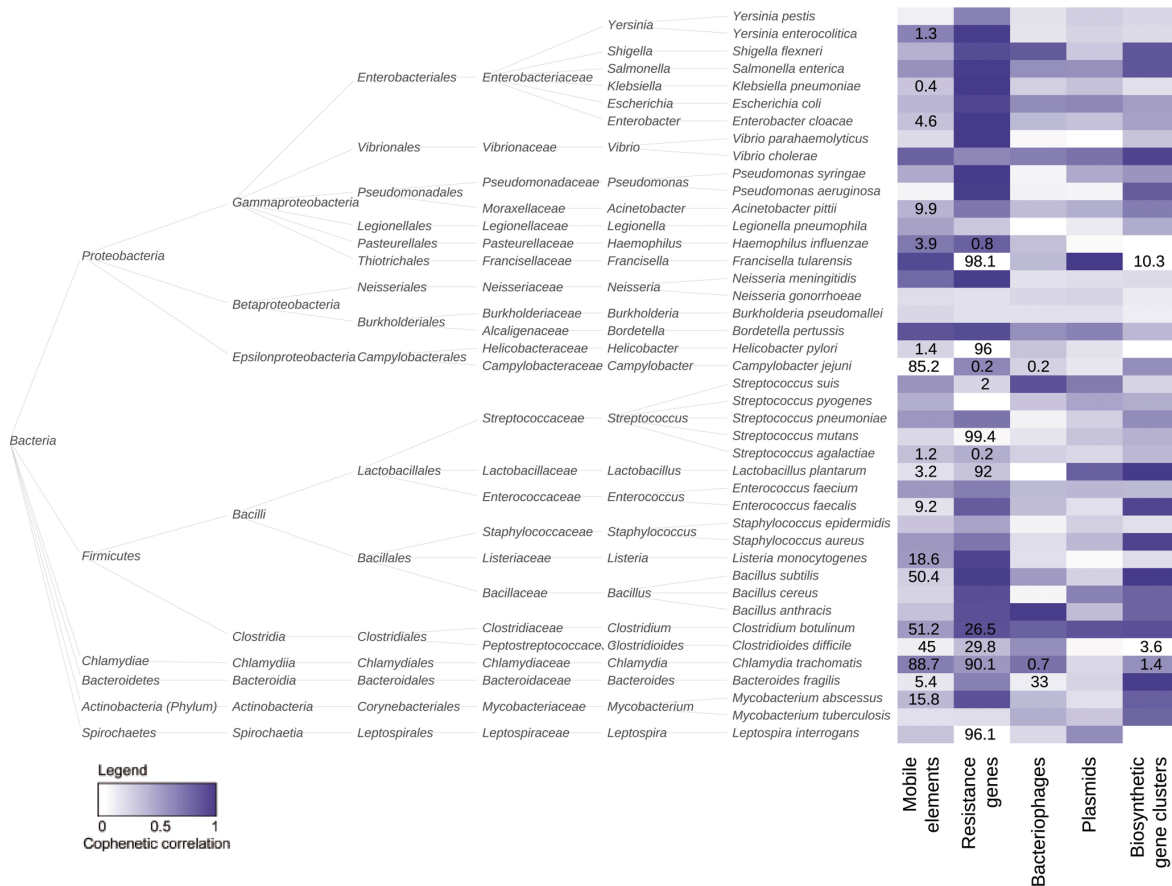


FIGURE 1.5 – Cophenetic distance between phenetic trees based on whole genome and filtered data sets for 42 bacterial species from RefSeq that included at least 100 genomes. Intensity of heatmap represents the cophenetic correlation as shown in the legend. Numbers in the heatmap are percentages of genomes with zero k-mers associated with relevant filtering data set.

phylogeny of the species.

The majority of the gammaproteobacteria had strong correlations with the ARG data set, especially species from *Klebsiella*, *Escherichia*, *Enterobacter*, *Vibrio*, *Pseudomonas*, and *Acinetobacter*. This could be related to the large number of intrinsic resistance determinants characterized in those species, especially the drug efflux systems [329]. Other studies have put into evidence the importance of bacteriophages and plasmids in the ongoing evolution of the *Vibrio* genus [160], as reflected in figure 1.5. *Shigella flexneri* is the Proteobacteria with the highest correlation with bacteriophages (0.83 CCC). Indeed, their O-antigens were often modified by serotype-converting bacteriophages [7; 375]. To further investigate this question, we used alignments to validate which bacteriophages used for filtering would be found in the 147 *Shigella* genomes. Interestingly, we found some specific prophage sequences that could delineate the clusters seen with clustering based only on phage k-mers (supplementary fig. 6, Supplementary Material online). Polysaccharides-related BGC, which encode capsular an-

tigens and O-antigens, could thus explain the high CCC of BGC for *S. flexneri* and *Vibrio cholerae* [63]. On the other hand, *E. coli* has several characterized BGC in the MIBiG database while showing moderate correlation with the whole genome (0.49 CCC) [259]. Comparison of clustering between whole genome and BGC of *E. coli* indicate that a portion of the population can be delineated by BGC while others seem unrelated (supplementary fig. 7, Supplementary Material online). The *Francisella tularensis* genome can contain over 100 insertion sequence genes [221], which could explain its high correlation with mobile elements. In opposition to most of the tested species, *F. tularensis* was also significantly correlated with plasmids. This high correlation could be related to a misannotated 100 kb plasmid that is in fact part of the *F. tularensis* genome (CP010448.1 which was replaced by CP010446.2). This large chromosomal region could indeed have boosted the impact of plasmids in the correlation observed, as it is integrated to the genome. It is important to consider that for most genomes in RefSeq, the plasmid sequences are found under a different accession number than the genome, therefore it is not considered in the clustering. In whole genome shotgun sequencing, plasmid sequences are generally included in the assemblies, thus plasmid filtering could prove useful to exclude these sequences from whole genome comparisons.

Six species from the *Firmicutes* phylum had correlation >0.70 CCC with ARG. *Bacillus anthracis*, *B. cereus* and *B. subtilis* were all above 0.85 CCC for ARG. This high correlation could originate from the chromosome-encoded β -lactamases harbored by the species [71; 117; 257]. The other members of the phylum, *Firmicutes* having good correlation with ARG, were *Listeria monocytogenes* 0.93 CCC, *Enterococcus faecalis* 0.82 CCC, and *S. pneumoniae* 0.70 CCC. The three *Bacillus* species also had CCC >0.70 for BGC. *Bacilli* are known to produce several types of secondary metabolites [341]. All the *Firmicutes* analyzed were below 0.55 CCC with the mobile elements data set. In *Firmicutes*, bacteriophages had best correlations with *Streptococcus suis* (0.87 CCC) and *B. anthracis* (0.98 CCC). The 500 *S. suis* genomes had an important number of k-mers associated with phages (18,642 in average), that along with the correlation, supported the idea that prophage sequences in the species are linked to the whole genome phylogeny. It was also shown in previous observations that remnants of phage sequences are distributed throughout *S. suis* genomes [379]. *Bacillus anthracis* had a high correlation with bacteriophages compared with the other *Bacillus* species, although shared k-mers from the filtering data set were not numerous (3,936 in average). The correlation could be related to four defective and conserved prophages harbored by the species as reported in Sozhamannan et al. (2006) [368]. In agreement with our results, they suggested that these prophages could be used as a chromosomal signature of the species. Bacteriophages could also be associated with ecological adaptation in *B. anthracis* [348].

Overall, the interpretation of the results represented in figure 1.5 supports our hypothesis that correlation between filtered genomes and complete genomes indicates a relationship between selected k-mers and a species. In many cases, we observed that a cophenetic correlation occur-

red in species where potentially mobile genetic elements were integrated in the genome. Thus, this methodology could potentially indicate integration and conservation of these elements in the genome of a particular species, or at least their phylotype dependence.

1.5 Conclusion

By comparing the k-mer composition of genomes, we were able to reconstruct the phenetic tree of large bacterial epidemiological genomics data sets, as we demonstrated with the *S. pneumoniae* and *P. aeruginosa* data sets. We also evaluated the accuracy of the methods on synthetic genome data sets by testing different parameters that influence this kind of analysis. The methodology is based on whole genome analysis rather than on a subset of core genes, which has been shown to introduce bias [358; 30]. The use of k-mers allows comparison of genomes based on characteristics that are either conserved or specific. We also applied the method to a data set of 2,429 bacterial genomes spanning the whole bacterial tree of life, without a selection of features such as conserved genes or ribosomal RNA. This approach makes Ray Surveyor an effective tool for scalable analyses in comparative genomics research, among other applications. Using k-mers to build phenetic trees could be used to easily position newly sequenced genomes in the microbial tree of life and infer classification or to determine which branches of the tree of life are not well represented in terms of genome sequences relative to internal taxa diversity.

Analysis of population structures can further be partitioned by filtering subsets of k-mers associated with gene categories or functions. Our results demonstrate that comparison of genomes based on specific subsets of k-mers can reveal their relationship at the population scale. Indeed, without being specific about the genetic determinants involved, the method allows easy determination of strain clusters with similar potential regarding the functions of the filtered data set, such as antibiotic resistance or HGT as shown in this study. A limitation of the filtering approach is that it involves the gathering of sequence data that adequately represents the diversity of the genes or functional category under study. For example, using only reference resistance genes instead of a large collection of orthologs, paralogs, and variants would underestimate the abundance of resistance genes in genomes containing variants of the reference gene. Still, some sequence types, such as bacteriophages or BGC, could be underrepresented in the databases used in this study. Such sequences, could have potentially resulted in more significant results, provided the availability of a more exhaustive and diverse sequence data set. As seen in figure 4 for the 2,429 bacterial genomes, some clusters of genomes show high bacteriophage signals in comparison to other regions of the heatmap. Indeed, of the 262 bacterial families included in the 2,429 genome analysis, 147 families had ≤ 100 k-mers associated with phage sequences, suggesting that some families could suffer from a lack of characterized phages in the database used for profiling (EBI). This issue should be alleviated by better filtering data sets as more sequences and better annotations become available in public databases.

Ray Surveyor is a powerful tool that allows the reconstruction and interpretation of the phenetic relationships underlying populations of bacterial species. By taking into account clinical or environmental context with the sequence filtering capabilities, this method could allow an

intuitive representation of population structures and the genomic features related to their differentiation or phenotype. It is thus a hypothesis-generating tool that could be applied to investigate the importance of specific gene categories not only in pathogens but also in environmental microbial communities and in the analysis of transcriptomic and metagenomic-based research.

1.6 Materials and Methods

Theoretical Background and Software Implementation

Ray Surveyor is built on top of the highly scalable Ray framework, which includes the Ray assembler and RayPlatform [34; 33]. It uses the message-passing interface (MPI) to scale analysis on supercomputers. However, depending on their size, datasets can be analyzed on smaller servers or personal computers. Components of the software include, among others, a sparse distributed hash table to store the k-mers on each computer across a cluster, as well as a graph coloring scheme that associates each k-mer vertex of the de Bruijn graph with its profiling datasets. Ray Surveyor is also based on the actor model [165]; each actor takes care of its own task such as reading and k-merizing input sequences, gathering k-mers into a store keeper, counting the k-mers, building the Gram matrix, etc. Supplemental Figure 7 provides further details on the actors' roles and their ways of communicating.

The first step of Ray Surveyor is to split the genome sequences into k-mers and build a graph of the pangenome. The k-mer length is set by the user. We recommend using a length between 21 and 61 nucleotides, usually 31 for the comparison of bacterial genomes. The workflow then proceeds with graph coloring, that assigns a virtual color for each k-mer according to the combination of genomes or functional datasets that carry it. The next step is to iterate over each k-mer and increment the count of shared k-mers between each pair of genomes of that color and store them in the Gram matrix. Formally, each pair of genome comparisons can be seen as a simple D_2 statistic [323; 406] with a binary count (presence/absence) of their k-mers. Since our counts are dichotomic, we can formally define the Ray Surveyor mechanics based on set theory.

Let $A_i = \{k_1, k_2, \dots, k_{l_A}\}$ be the set of all the k-mers of genome i , and similarly $B_j = \{k_1, k_2, \dots, k_{l_B}\}$ the set of all the k-mers of genome j . Then, the Gram matrix (K) is defined such that $k_{i,j} = |A_i \cap B_j|$. Let $Z = \{z_1, z_2, \dots, z_m\}$ be m filtering datasets and $Y = \bigcup_{n=1}^m Z_m$ their union. To filter in (include only) the k-mer set Y , $k_{i,j} = |A_i \cap B_j \cap Y|$ and to filter out (exclude) the k-mer set Y , then $k_{i,j} = |(A_i \cap B_j) \setminus Y|$. The resulting matrix K is then normalized to have values in the range $[0, 1]$, with the diagonal entries equal to 1. Consequently, the entries of the normalized matrix K' are given by $k'_{i,j} = \frac{k_{i,j}}{\sqrt{k_{i,i} * k_{j,j}}}$. However, when filtering is used, we recommend division of the entries $k_{i,j}$ by the $k_{i,i}$ and $k_{j,j}$ of the full k-mer matrix, rather than the filtered version. The reason is that the diagonal of the filtered matrix no longer represents the total number of k-mers per genome, but only the number of filtered k-mers, a subset of the genome. This renders the matrices more comparable, as they are all normalized with respect to the same total k-mer content.

After normalization, the matrix is transformed into a distance matrix with a chosen metric. We focused our experiments on four metrics that are the cosine, correlation, Euclidean and

Canberra. Below, we formally define the distance formulae by using u and v and the normalized vectors of shared k-mers between a genome and all the other genomes in the population. For instance, the entry $d_{1,2}$ in the distance matrix D , would be defined as $d_{1,2} = 1 - \frac{k'_1 \cdot k'_2}{\|k'_1\|_2 \|k'_2\|_2}$ for the cosine distance metric. With the vectors $u = k'_1$ and $v = k'_2$, here are the formula of the four distance metrics tested in our study :

— cosine :

$$1 - \frac{u \cdot v}{\|u\|_2 \|v\|_2} \quad (1.1)$$

— correlation :

$$1 - \frac{(u - \bar{u}) \cdot (v - \bar{v})}{\|(u - \bar{u})\|_2 \|(v - \bar{v})\|_2} \quad (1.2)$$

— Euclidean

$$\|u - v\|_2 \quad (1.3)$$

— Canberra :

$$\sum_i \frac{|u_i - v_i|}{|u_i| + |v_i|}. \quad (1.4)$$

An important limitation of the cosine and correlation distances is that they cannot be evaluated if one of the vectors only contain zeros. This means that if a genome does not share any k-mer with all the other genomes, the two metrics will fail with an undefined behavior due to the division by zero (from $\|u\|_2$ or $\|v\|_2$). This may also happen when we filter the comparison with a functional dataset and there is one genome that doesn't harbor any k-mer from it. The two other metrics (Euclidean and Canberra) are robust to those outliers without shared k-mers but their results are still influenced by them. Hence, species with a large proportion of genomes containing no k-mer of the filtering dataset should not be interpreted with this methodology. Undefined distances with cosine and correlation metrics were set to zero in our experiments. For this reason, in the manuscript, figures showing cophenetic distance of filtered datasets used the Canberra distance.

The matrix computation in Ray Surveyor uses the SciPy python package [188]. Computation of distance metrics can also be performed with R software. Moreover, the Ray Surveyor scripts allow computation of a Newick tree from the distance matrix either with the neighbor-joining or UPGMA method (unweighted pair group method with arithmetic mean) based on the scikit-bio and BioPython packages [67].

1.7 Phenetic and Phylogenetic Analysis

Simulated datasets

Simulated trees with three different average branch lengths (0.001, 0.005, 0.01) were randomly produced to represent different evolutionary distances of 100 genomes [213; 150]. For each of the 3 average branch lengths, we generated 10 trees to evaluate reproducibility. Sequence alignments of 1,000,000 sites were derived from the 100 genomes' trees based on a simple nucleotide model (equal equilibrium frequencies and equal mutation rates) from the Pyvolve python package [370]. The sequences obtained from the gapless alignments were used for subsequent Ray Surveyor analyses. The four distance metrics (Euclidean, cosine, correlation, Canberra) were tested in our simulation to transform Ray Surveyor's similarity matrix into a distance matrix. We also tested ten different k-mer lengths - ranging from 11 to 101 with an increment of 10 - to evaluate their performance. To ensure the validity of our tree and sequence models, an alignment-based phylogeny with the FastTree NT-GTR model [310] was made for all the trees. The alignment-based phylogenies were also compared with the reference phylogeny using the same methods as for Ray Surveyor clusters (phenetic trees) or neighbor-joining trees. Two evaluations were made to test how well our method would replicate the reference simulated trees. First, the simulated tree distance matrices were compared to Ray Surveyor's distance matrices with the cophenetic correlation coefficient (CCC) using the ape [294] and dendextend [129] R packages. CCC indicates how similar the pairwise distances are between two dendrograms obtained by hierarchical clustering from the distance matrix. These dendrograms are referred to as phenetic trees throughout the manuscript. Secondly, the topology of the trees was compared with the Robinson-Foulds (RF) metric with the ETE3 python package [171]. RF counts the minimal number of branch operations required to change one tree into the other. The average nucleotide identity (ANI) was also computed for all the simulated alignment sequences. The ANI statistics for all the trees are reported in Supplementary Table 1.

Real prokaryotic genome datasets

The phylogenies and metadata for the Croucher et al. *S. pneumoniae* dataset and the Kos et al. *P. aeruginosa* dataset were obtained from the authors [75; 76; 211; 96]. The phylogeny of the Hilty et al. *S. pneumoniae* dataset was obtained using 602 conserved genes aligned with MAFFT v7.221 [199]. A maximum likelihood phylogeny was performed on the 602 concatenated genes with RAxML version 8.1.20 [374]. In order to compare phylogenetic trees with the clusters of Ray Surveyor, the trees were converted from their Newick format into a cophenetic distance matrix using the R package : ape [294]. Hierarchical clustering was performed using the UPGMA (average) method. The 2429 bacteria genome phylogenetic tree was based on the 16S rRNA gene and taxonomical annotation was based on the established NCBI taxonomy. Initially, 2429 bacterial genomes were obtained from NCBI (See Supplementary Table 2 for a list). To build the phylogeny of the bacterial tree of life, the 16S rRNA gene sequences were

extracted from each genome. Then, the 2429 16S rRNA genes were aligned using MAFFT v7.221 [199] and a maximum likelihood phylogeny was produced with RAxML version 8.1.20 [374]. Cophenetic correlation coefficient and Fowlkes-Mallows index were calculated with the dendextend R package [129]. Ray Surveyor was run with a k-mer length of 31 to keep a high stringency in the coloring of the graph [34]. The 2429 bacterial genomes similarity matrix was produced with Ray Surveyor on a computer cluster using 4 nodes of 48 cores with 256GB of RAM for a total compute time of less than 6 hours.

Source of tools and datasets

Ray Surveyor is freely available under the GPLv3 license at <https://github.com/zorino/ray>. A tutorial on how to run an analysis is available at <https://github.com/zorino/raysurveyor-tutorial>. The 2429 bacterial genomes were downloaded from the NCBI GenBank (<ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/>) in September 2015. Only the sequences marked either as a representative or a reference genome in the assembly reports were selected. The goal was to compute phylogenetic trees and clustering from a limited number of genomes that represented a broad taxonomical overview of the domain Bacteria. Since the NCBI GenBank genome database has an inherent bias towards certain taxa [383], such as clinically relevant pathogens, it allowed us to discard a large number of similar genomes. The total number of nucleotides analyzed in this dataset was 11.4 billion with an average of 3.9 million per genome. The targeted analyses of *Streptococcus pneumoniae* and *Pseudomonas aeruginosa* were extracted from the literature [75; 211] and downloaded from NCBI GenBank or ENA. The datasets of resistance genes and mobile elements were taken from the MERGEM database (<http://mergem.genome.ulaval.ca>) [322], the plasmids were taken from the NCBI Plasmids collection in June 2015, the bacteriophage from the EBI collection in June 2015 (<http://www.ebi.ac.uk/genomes/phage.html>) and the biosynthetic gene clusters from the MIBIG v1.0 database [259].

1.8 Supplementary Material

Supplementary data are available at Molecular Biology and Evolution online.

1.9 Author Contributions

MD and FR performed bioinformatics analyses. MD and SB programmed the Ray Surveyor software. MD, SB and FL designed algorithms. MD, FR, AC, PHR and JC interpreted biological results. MD, FR, AC and JC contributed to the preparation of the manuscript. All authors critically reviewed the manuscript.

1.10 Acknowledgments

This study was financed by the Canada Research Chair in medical genomics (JC). FR was supported by a Mitacs post-doctoral fellowship. MD was supported by the Fonds de recherche du Québec - Santé. The authors thank Pier-Luc Plante, Alexandre Drouin, Pascal Belleau and Maurice Boissinot for their comments. Computations were performed under the auspices of Calcul Québec and Compute Canada. The operations of Compute Canada are funded by the Canada Foundation for Innovation (CFI), the National Science and Engineering Research Council (NSERC), NanoQuébec and the Fonds Québécois de Recherche sur la Nature et les Technologies (FQRNT).

Chapitre 2

Base de données protéique flexible et efficace basés sur kAAmer

Titre original

Fast protein databases as a service using kAAmer

Journal

Première version du manuscrit en pré-impression dans bioRxiv (<https://www.biorxiv.org/content/10.1101/2020.04.01.019984v1>).

Présentement en révision dans Scientific Reports (<https://www.nature.com/srep/>).

Auteurs

Maxime Déraspe, Sébastien Boisvert, Paul H. Roy, François Laviolette, Jacques Corbeil

2.1 Résumé

En génomique, l'identification des protéines au sein de séquences d'ADN est l'une des tâches les plus intensives au niveau du calcul informatique. Ce sont généralement des logiciels d'alignements qui exécutent ses tâches et ces derniers n'incorporent pas d'informations riches sur les protéines qui sont souvent nécessaires aux annotations de génomes. Le présent article introduit kAAmer, un engin de bases de données de protéines basé sur les k -mers d'acides aminés. kAAmer supporte l'identification rapide de protéines en plus de fournir des annotations détaillées sur celles-ci. Il se veut un outil rapide et flexible conçu pour héberger des bases de données protéiques en tant que services que ce soit dans l'infonuagique ou sur du matériel informatique dédié.

2.2 Abstract

Identification of proteins is one of the most computationally intensive steps in genomics studies. It usually relies on aligners that don't accommodate rich information on proteins and require additional pipelining steps for protein identification. We introduce kAAmer, a protein database engine based on amino-acid k-mers, that supports fast identification of proteins with complementary annotations and can be hosted and queried remotely.

2.3 Main

One fundamental task in genomics is the identification and annotation of DNA coding regions that translate into proteins via a genetic code. Protein databases increase in size as new variants, orthologous and paralogous genes are being sequenced. This is particularly true within the microbial world where bacterial proteomes' diversity follows their rapid evolution. For instance, UniProtKB (Swiss-Prot / TrEMBL) [387] and NCBI RefSeq [288] contain over 100 million bacterial proteins and that number continues to grow rapidly.

Identification of proteins often relies on accurate, but slow, alignment software such as BLAST or hidden Markov model (HMM) profiles [46; 102]. Although other approaches (such as DIAMOND [43]) have considerably improved the speed of searching proteins in large datasets, from a database standpoint much can be done to offer a more versatile experience. One advantage would be to expose the database as a permanent service making use of computational resources for increased performance (i.g. memory mapping) and leveraging the cloud for remote analyses via a Web API. Another aspect would be to extend the result set with comprehensive information on protein targets to facilitate subsequent genomics and metagenomics analysis pipelines.

Alignment software usually relies on a seed-and-extend pattern using an index (two-way indexing in DIAMOND) to make local alignments between query and target sequences. However, there is a plethora of research techniques to bypass the computational cost of alignment. Alignment-free sequence analyses usually adopt k-mers (overlapping subsequences of length k) as the main element of quantification. They are extensively used in DNA sequence analyses ranging from genome assemblies [306] to genotyping variants [178], as well as genomics and metagenomics classification [34; 421; 289]. In the present study, we introduce kAAmer, a fast and comprehensive protein database engine that was named after the usage of amino acid k-mers which differs from the usual nucleic acid k-mers. We demonstrate the usefulness and efficiency of our approach in protein identification from a large dataset and antibiotic resistance gene identification from a pan-resistant bacterial genome.

The database engine of kAAmer is based on log-structured merge-tree (LSM-tree) Key-Value (KV) stores [290]. LSM-trees are used in data-intensive operations such as web indexing [52; 136], social networking [109] and online gaming [80; 246]. kAAmer uses Badger [92], an efficient implementation in Golang¹ of a WiscKey KV (key-value) store [246]. WiscKey's LSM-tree design is optimized for solid state drives (SSD) and separates keys from values to minimize disk I/O amplification. Disk I/O amplification is typical of LSM-trees due to its vertical design in which keys and values need to be read and rewritten in multiple levels of the tree. Therefore, kAAmer will obtain peak performance with modern hardware such as NVMe²

1. Go programming language (<https://golang.org/>)
2. Non-Volatile Memory Express

SSDs. Furthermore, traditional block devices such as SATA solid-state drives that offer good throughput in input/output (I/O) operations per second (IOPS) should effectively accommodate use cases where many queries are sent simultaneously. A kAAmer database includes three KV stores (see Figure 2.1) : one to provide the information on proteins (protein store) and two to enable the search functionalities (k-mer store and combination store). The k-mer store contains all the 7-mers found in the sequence dataset and the keys to the combination store, which uniquely serves the combination of proteins held by k-mers. The fixed k-mer size at 7 was chosen to fit on 4 bytes and keep a manageable database size while offering good specificity over protein targets. The k-merized design of a kAAmer database provides an interesting simplicity for the search tasks which will give an exact match count of all 7-mers between a protein query and all targets from a protein database. The result sets using this strategy are not guaranteed to return the same homologous targets that would be obtained with alignment or HMM search and is therefore less suitable for distant homology retrieval ($< 50\%$ identity). Nonetheless, kAAmer also supports alignment on the result set without sacrificing too much speed as shown in Figure 1B. The main drawback of a kAAmer database (in the version at the time of writing : 0.4) is the disk space and time required to build a database that is greater than its benchmarked competitors, although it compares favorably to Ghostz [134] for these parameters.

In order to test the efficiency of our database search engine, we used all (114,830,954) the non-fragmented proteins of the UniprotKB (Swiss-Prot / TrEMBL) bacterial proteins dataset (release 2019_08). Sixteen different protein query datasets were randomly and uniquely chosen from the original database, with size ranging from 1 protein to 10,000 proteins. We added the kAAmer search in k-mer match mode (without alignment ; named “kaamer-kmatch”) for comparison purposes. We also corrected the kAAmer alignment mode (“kaamer-aln”) in “kaamer-aln+opendb”, by adding the time it took to open the database before running the queries (230 seconds). However, kAAmer’s purpose is to be used as a persistent service so the database opening time becomes insignificant the more you query the database. The four software included in the benchmark are Blastp (v2.9.0+) [46], Ghostz (v1.0.2) [134], Diamond (v0.9.25) [43] and kAAmer (v0.4). Figure 1B illustrates the wallclock times of the alignment software in comparison with kAAmer for protein homology searches. See the Methods section for the hardware used in the benchmarks. We observe that with the larger query datasets (10,000 proteins), kAAmer in alignment mode completes its search and alignments in just 3 minutes 2 seconds (and 2 minutes 26 seconds without alignment). In comparison, Diamond, the second fastest aligner, achieved the 10,000 query task in 11 minutes 57 seconds. Thus, at the maximal benchmarked query size, kAAmer shows an increase in speed of almost 4x (3.93x). When the search incorporates fewer protein queries the gain of kAAmer is more substantial (up to 82x with only 10 protein queries) because Diamond and its double indexing is optimized to perform better when the number of queries increases. It is worth mentioning that when correcting for the database opening (square symbol in Figure 1B.), the kAAmer

gain in speed drops and it only surpasses Diamond when there are over 4000 protein queries. However, as stated earlier, kAAmer is rather suited to act as a permanent and flexible database service that will store structured protein information and offer a quick homology search over that protein database. Also, with sufficient random access memory (RAM), data is going to be cached by the operating system (OS) which will increase the performance of kAAmer. For the other benchmarked software, Ghostz took over 33 minutes to realise the task with the 10,000 queries, which is 11 times slower than kAAmer. For Blastp, we stopped the benchmark at 2,000 protein queries since it was already taking over 7 hours to complete the task (at least 700x slower than kAAmer).

In order to accommodate real-use cases we built relevant kAAmer databases and investigated their usage in typical bacterial genomics analyses. It should be noted that annotation of genomes and gene identification rely heavily on the quality of the underlying database. What kAAmer has to offer is the inclusion of the protein information within the database combined with an efficient search functionality to facilitate downstream analyses. Therefore, we also provide utility scripts to demonstrate these use cases. The first use case was to identify antibiotic resistance genes (ARG) in a bacterial genome and test its accuracy related to other ARG finder software. For ARG identification we used the NCBI Bacterial Antimicrobial Resistance Reference Gene Database (v2020-01-06.1) [112] and compared the kAAmer results with the ResFinder (v3.2 and database 2019-10-01) [435] and CARD (v5.1.0) [5] software and database. The query genome is a pan-resistant *Pseudomonas aeruginosa* strain E6130952 [425]. Table 2.1 shows the results of the ARG identification within the query genome by the three software / databases tested. For the majority of antibiotic classes, the results are in agreement between the three databases. Interestingly, three aminoglycoside genes (*aac(6')-II*, *ant(2'')-Ia* and *aacA8*) were only found with kAAmer (NCBI-ARG) and ResFinder. On the other hand, several more antibiotic efflux systems are annotated in CARD and the number of identified efflux proteins in E6130952 goes up to 36 while only 3 were reported by kAAmer (NCBI-ARG) and none by ResFinder. Also 2 genes associated with resistance to peptide antibiotics (*arnA*, *basS*) and 2 other (*soxR*, *carA*) associated with multiple antibiotic classes were only reported by CARD. Other tested use cases include genome annotation and metagenome profiling as shown in the Methods section.

In summary, kAAmer introduces a fast and flexible protein database engine to accommodate different genomics analyses use cases. It can be hosted on-premise or in the cloud and be queried remotely via an HTTP API.

2.4 Methods

2.4.1 Design of kAAmer

kAAmer design was influenced by our requirement that protein databases would be permanently hosted (on premise or in the cloud), queried remotely and would have room to scale as sequence databases grow in size. It also needed to be multithreaded for protein searches and would support alignment for more accurate remote homology findings. We opted for a Key-Value store engine that would reside on disk and be optimized for SSDs. We used the Go programming language for its versatility and efficiency. The Key-Value stores use the Badger [92] engine and protein annotations are encoded using Protocol Buffers [145].

2.4.2 Database building

kAAmer is first used to build a database in which all amino acid k-mers are associated with proteins in which they are found. It consists of three KV stores to hold the database information (k-mer store, combination store and protein store). The first KV store (k-mer store) keeps the association of every k-mer (key) with a hash value (key length : 8 bytes) that is the entry to the combination store. The k-mer size is fixed at 7 amino acids to fit k-mer keys onto 32 bits (4 bytes) and thus maintain a manageable final database size while keeping a k-mer size long enough for specificity. The second KV store (combination store) is used to hold all the unique sets of protein identifiers. The method used to build this store can relate to the flyweight design pattern, the hash consign technique, and the coloured de Bruijn graphs [178; 34]. Indeed, hash values are reused to access identical objects and therefore minimize memory usage. The set of protein identifiers are the keys to the last store (protein store) which contains the protein information found in the raw annotation file. The raw input file can be either in the EMBL format, GenBank format, TSV format or in FASTA format.

2.4.3 Database querying

Once we have a database, we expose it with the kAAmer server that listens over HTTP for incoming requests. The benefits of using such a service are two fold. First, the database is opened once and is memory mapped to increase the performance of protein searches. Second, the kAAmer server can be hosted virtually anywhere, in the cloud for instance, and be queried remotely by the kAAmer client. Note that it is preferable that the latency (time required for a message to be transported over HTTP) between the server and client be as low as possible. kAAmer supports protein query and translated DNA query from FASTA input as well as short reads sequences (like Illumina) in FASTQ format.

2.4.4 Protein benchmark software

To build the benchmark on the UniProtKB bacterial proteins database, we randomly and uniquely extracted multiple sets of sequences, with the number of sequences ranging from 1 to 10 thousand. Each set of sequences was in its own FASTA file to be queried with the different alignment software included in the benchmark. The benchmark for all four software (Blastp (v2.9.0+), Ghostz (v1.0.2), Diamond (v0.9.25) and kAAmer (v0.4)) was run on nodes geared with 32 cores (Intel(R) Xeon(R) CPU E5-2667), 120 GB of RAM and with a SATA III connected SSD. The maximum number of results for each query was set to 10 and no threshold was provided. All software were run with default parameters, except for the number of threads set to 32 and the maximum number of results per query at 10.

2.4.5 Other kAAmer use cases

Apart from the antibiotic resistance gene (ARG) identification use case, we also provide two demonstrations of kAAmer usage in bacterial genome annotation and metagenome profiling. The use cases are documented at <https://github.com/zorino/kaamer-demo> and a Python script is provided for each one of the analysis. For the genome annotation, we used the chromosomal sequence of the same *Pseudomonas aeruginosa* strain (E6130952) as in the antibiotic resistance genes identification. The kAAmer database that was used for the homology detection is a subset of RefSeq from the *Pseudomonadaceae* family which is available from the kAAmer repository (see Data availability) along with other Bacterial family databases. Essentially the genome annotation script parses the kAAmer results and produces a GFF (General Feature Format) annotation file giving some threshold on the protein homology. The other use case is the profiling of a metagenome based on the MGnify database of the human gut [265]. MGnify includes protein annotations from gene ontology, enzyme commission and kegg pathways, among others. The metagenome profiling script will parse the results and produce a summary file by annotation that counts the presence and abundance of each feature.

2.5 Data availability

We have built a repository where one can download prebuilt kAamer database versions of common protein datasets useful in bacterial genomics and metagenomics. The repository is available at <https://kaamer.genome.ulaval.ca/kaamer-repo/> and includes datasets from the NCBI, the EBI and other popular data sources.

2.6 Code availability

The code is available at <https://github.com/zorino/kaamer> under the Apache 2 license and the documentation can be found at <https://zorino.github.io/kaamer/>.

2.7 Acknowledgements

This study was financed by the Canada Research Chair in medical genomics (JC). MD was supported by the Fonds de recherche du Québec - Santé (#32279). The authors thank Juan-Manuel Dominguez and Charles Burdet for their comments. Computations were performed under the auspices of Calcul Québec and Compute Canada. The operations of Compute Canada are funded by the Canada Foundation for Innovation (CFI), the National Science and Engineering Research Council (NSERC), NanoQuébec and the Fonds Québécois de Recherche sur la Nature et les Technologies (FQRNT).

2.8 Declaration

The authors declare no competing interests.

2.9 Figures and tables

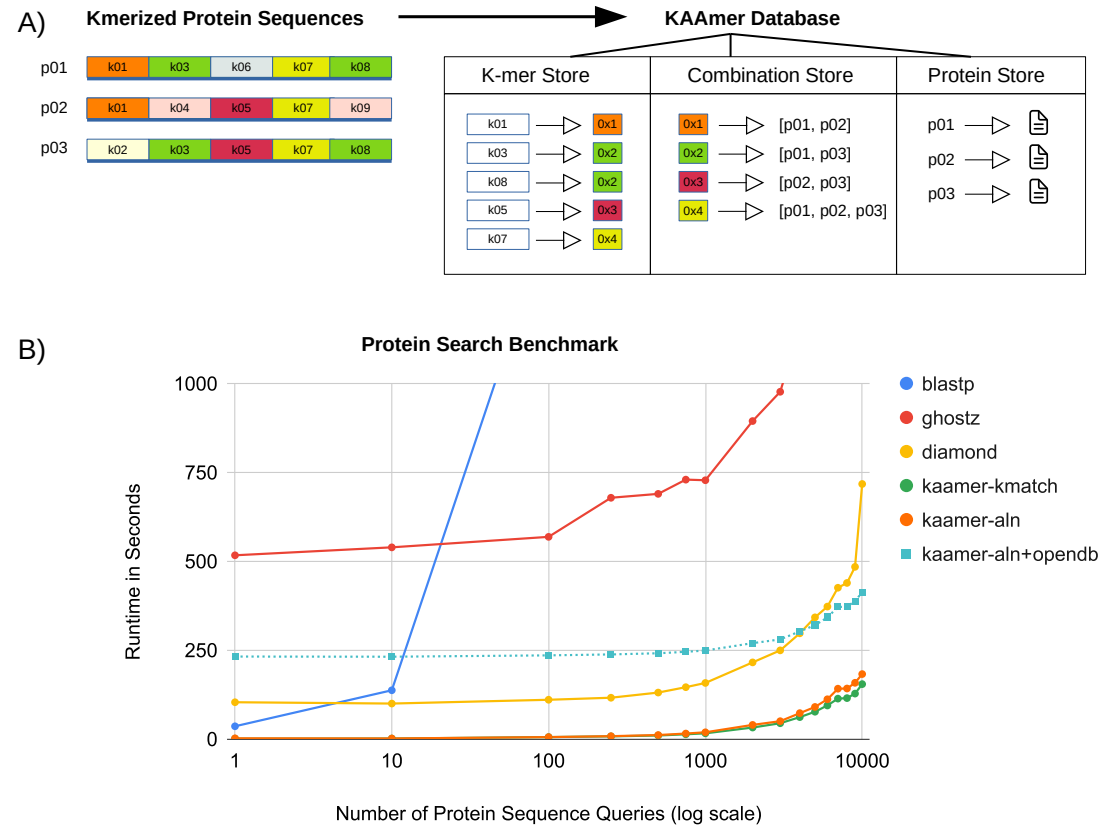


FIGURE 2.1 – A) Design of a kAAmer database. Three key-value stores are created within a database (K-mer Store, Combination Store, Protein Store). Colours indicate the combination (hash) value that are reused in the combination store. Proteins are numbered (p01, p02, p03) and k-mers are numbered (k01,k02,...,k08). B) Protein search benchmark. Software include blastp (v2.9.0+), ghostz (v1.0.2), diamond (v0.9.25) and kAAmer (v0.4) with and without alignment.

TABLEAU 2.1 – Report of the antibiotic resistance genes identification within the pan-resistant *Pseudomonas aeruginosa* E6130952 strain from kAAmer+NCBI-arg, ResFinder and CARD databases.

Resistance Gene	Antibiotic Class	kAAmer+NCBI-ARG	ResFinder	CARD
aac(6')-II (aacA7)	amikacin/kanamycin/tobramycin	3	3	0
ant(2'')-Ia	gentamicin/kanamycin/tobramycin	1	1	0
aacA8	aminoglycoside	1	1 (aac(6')-31)	0
aph(3')-IIB	kanamycin	1	1	1
aadA6	streptomycin	2	2	2
blaOXA-2, blaOXA-488	beta-lactam	2	2	2
blaPDC-35	cephalosporin	1	1 (blaPAO)	1 (blaPDC-2)
fosA	fosfomicin	1	1	1
catB7	chloramphenicol	1	1	1
sul1	sulfonamide	3	3	3
mexA, mexE, mexX	efflux	3	0	2 (no mexX)
other efflux system	efflux	0	0	34
arnA, basS	peptide antibiotic	0	0	2
soxR, carA	multiple antibiotic class	0	0	2
Total	13	19	16	51

Chapitre 3

Analyses multi-études du microbiome intestinal en relation avec l'obésité

Titre original

Cross-study analyses of gut microbiomes from healthy and obese individuals - Deraspe2020

Journal

Soumis à PLOS ONE (<https://journals.plos.org/plosone/>).

Auteurs

Maxime Deraspe, Charles Burdet, Juan Manuel Dominguez, Paul H. Roy, François Laviolette, Jacques Corbeil

3.1 Résumé

Un quantité croissante d'études métagénomiques sont menées dans le but de mieux comprendre les modifications du microbiote intestinal en relation avec différentes conditions de santé. Des résultats significatifs ont été obtenus pour plusieurs maladies telles que les cirrhoses, les cancers colorectaux et les maladies inflammatoires de l'intestin. La relation entre le microbiote et la condition de l'obésité se démarque par les résultats contradictoires qui ont été rapportés dans la littérature. Nous avons donc construit et analysé un ensemble de données de sujets sains et obèses, provenant de différentes études, pour évaluer les changements majeurs dans la composition taxonomique et fonctionnelle de leurs métagénomes. Nos résultats suggèrent que les sujets en surpoids et les sujets normaux n'ont pas de fortes dissemblances dans la composition de leurs métagénomes. Des différences significatives sont cependant observées en comparant les individus obèses et non obèses dans leurs profils fonctionnels et taxonomiques. Dans cette étude, nous rapportons les changements les plus significatifs que nous avons observés et discutons de leur implication potentielle dans la condition d'obésité.

3.2 Abstract

With the advent of metagenomics, many large studies have been conducted with the quest of better understanding gut microbiota changes in relation to varying health conditions. Significant findings have been made for diseases such as cirrhosis, colorectal cancers, inflammatory bowel diseases and others, yet one that stands out is obesity for which conflicting results have been reported in the literature. Here, we built and analyzed a cross-study dataset of healthy and obese individuals looking for major changes in the the taxonomic and functional composition of their metagenomes. Our results suggest that the overweight and normal subjects have no strong dissimilarity in their metagenomes composition. Significant differences were observed when comparing the obese and the non-obese individuals in their functional and taxonomic profiles. In this study, we report the most significant changes that we observed and discuss their potential implication in the obesity condition.

3.3 Background

Obesity is a major epidemic with an economic cost to society exceeds 150 billion dollars in health resource allocation in the United State alone [173]. Since the 1980s in more than 70 countries, the prevalence of obesity has more than doubled, while increasing in most other countries [70]. At least three factors have a major impact on the body mass index (BMI) increases : first, the presence of trans fatty acids in food consumption ; second, the high fructose levels in soda and fruit juices ; and third, physical inactivity [173]. More importantly, increase in BMI is linked to several chronic diseases including cardiovascular diseases [385], diabetes mellitus (especially type 2) [6], chronic kidney disease [324], numerous cancers [308], and musculoskeletal disorders [291].

The gut microbiome is also, by all accounts, impacted by our lifestyle (diet, exercise, drug use) and is even considered, rightly so, as a human microbial organ [25]. Numerous studies relate dysbiosis of the gut microbiome with various pathological conditions such as obesity [223], type 2 diabetes [316; 194], inflammatory bowel disease [277; 275], colorectal cancer [436; 115; 404] or even mental disorders [49]. Pasolli and collaborators examined several diseases associated with microbiome datasets which they reanalyzed with machine learning aiming to provide an evaluation of the classification of ill and healthy individuals based on the gut metagenomics data [297]. In all the assessed conditions (cirrhosis, colorectal, IBD, T2D and obesity [223]), obesity remains the worst predictive results with an AUC not surpassing 0.66. This same group also succeeded in improving their results by performing cross-study sample inclusion with additional control samples. However, this specific approach was not performed for the obesity cohort.

Several controlled studies have examined the relationship between obesity and the gut microbiota composition [228; 229; 394; 392; 349]. A general observation was the the ratio of the two taxonomic phyla *Bacteroidetes* and *Firmicutes* between obese and non-obese human or animal (mouse) [228; 229; 394; 100; 392; 349; 223]. The observation was not consistently reproduced as no significant relationship was found for the *Bacteroidetes* phylum in Duncan *et al.*, and in Schwartz *et al.* whereas the *Firmicutes* to *Bacteroidetes* ratio was increased in favour of *Bacteroidetes* in obese subjects contrary to what was found in two studies by Ley *et al.* and the 2006 study of Turnbaugh *et al.* where the ratio was inverted.

In our study, we addressed and investigated the creation of a cross-study human gut microbiome dataset and the evaluation of their composition when analyzing individuals' BMI. We used the MGnify protein database as the core resource for all our analyses [265]. We also sought to classify of obese and non-obese subjects based on each protein feature annotation of the gut metagenomes by evaluating different machine learning algorithms on these features. Note that for convenience, the words healthy and normal are used interchangeably in this manuscript and it refers to a BMI between 18.5 and 30 (healthy and overweight based on

the CDC categories), as well as for the word obese that represent a BMI ≥ 30 , without any distinction of the obesity subtypes if not mentioned otherwise.

3.4 Methods

3.4.1 Cross-study samples selection

The dataset was created by collating all the metadata available from the curatedMetagenomic-Data R package [296] and by selecting samples that respected the following selection criteria : subjects aged between 18 and 65 years old were considered healthy according to each study's procedure (except for overweight), and there was no indication of drugs used to treat any underlying conditions. Their BMI was between 18.5 and 30, subjects were adults. Similarly, the obese samples were chosen with these criteria except that their BMI exceeded 30. Subjects were chosen from different countries (Netherlands 378, Denmark 109, Madagascar 83, Spain 40, USA 19, Peru 11) to ensure a reasonable geographic distribution that would extend our analysis to different ethnic populations and therefore include different diet and genetics. Altogether, the gut metagenomes originate from four different studies with one unpublished (PasolliE_2018) [277; 284; 345]. The total number of samples summed up to 640 of which 221 were male subjects compared to 270 females and 149 of unknown gender. As for the age of the subjects, 58.9% were between 18 and 30, while the remainder were between 31 and 65.

3.4.2 Definition of overweight and obesity

The official CDC¹ body fatness categories associated to BMI are < 18.5 for underweight, 18.5 to < 25 for normal, 25 to < 30 for overweight and 30 to < 35 , 35 to < 40 and ≥ 40 for class 1,2 and 3 obesity, respectively. In our reported results we have combined the obesity 2 and 3 levels due to the small sample size as only 3.3% of all samples have a BMI ≥ 35 .

3.4.3 Metagenomics data preparation

The whole genome sequencing (WGS) raw sequence reads (FASTQ files) of the publicly available gut metagenomes from included subjects were downloaded from the NCBI² sequence read archive (SRA). Quality filtering on the reads was performed using fastp[58], with a quality Phred score of 30. In order to normalize the depth of all the samples, the filtered reads were subsampled at 10 millions reads. We chose the Unified Human Gastrointestinal Protein (UHGP) catalogue clustered at 90% identity (UGGP-90) from MGnify databases [265] for the metagenome annotation using kAamer³ [86] for protein identification. MGnify includes annotations from eggNOG [172], the enzyme commission database [21], and KEGG pathways

1. Centers for Disease Control and Prevention : <https://www.cdc.gov/obesity/adult/defining.html>

2. National Center for Biotechnology Information

3. <https://github.com/zorino/kaamer>

[191], among others. Extraction of the significant hits was performed with the analysis scripts provided with kAAmer. We also assembled the subsampled reads into contigs using megahit [231] and did a followup analysis with Ray Surveyor [90] to compare the metagenomes DNA k-mer content.

3.4.4 Statistical analyses

We compared the subjects considered as normal for their body fatness condition ($BMI < 30$) with those classified as obese ($BMI \geq 30$). Statistical analyses were performed in Python using the *scikit-bio*, *statsmodels* and *scipy* libraries [350; 188]. We compared the distribution of each protein feature (Taxa, Gene Ontology, Enzyme Commission, COG, KEGG pathways and modules) using the non parametric Wilcoxon - Mann & Whitney test, and p-values were corrected using the Benjamini-Hochberg procedure as implemented in the *statsmodels* Python package. Alpha diversity was computed using the Shannon index with the *scikit-bio*⁴ Python package and compared between groups using the non parametric Wilcoxon - Mann & Whitney test. Beta diversity computation was also achieved using *scikit-bio* with the Bray-Curtis dissimilarity and compared between cohorts with the analysis of similarities (ANOSIM) test.

3.4.5 Machine learning in the comparison of the microbiomes

We used the utility functions from the kAAmer analyses scripts to compute machine learning classifiers based on the protein feature annotations provided in the MGnify database (taxa, Gene Ontology, Enzyme Commission, KEGG pathways and modules, and Clusters of Orthologous Groups (COG)) [265]. The machine learning algorithms tested included gradient boosting methods (xgboost [59], lightgbm [200], catboost [311]), ensemble method (random decision forests [168]) and the support vector machine (SVM [161]). The number of subjects in both groups was unbalanced with a ratio of approximately 6.8 in favour of the normal cohort - 558 normal individuals against 82 obese individuals - and the training was performed accordingly. All hyperparameter searches were performed using the Optuna framework with a 10 fold cross-validation [4]. Similarly, the score metrics were evaluated with a 10 fold stratified cross-validation. The cross-validation and score evaluations were achieved using the scikit-learn package [301]. Feature importance from the models was also extracted using the ELI5's Python package⁵.

4. <http://scikit-bio.org/>

5. <https://github.com/TeamHG-Memex/eli5>

3.5 Results

3.5.1 K-mer analyses of the metagenomes

Metagenomes comparison based on their complete genomics content is a complex task due to the vast diversity of microbes found in individuals and the subsequent very high number of genetic sequences obtained by NGS methods. Ray Surveyor allows the direct comparison of metagenomes based on their k -mers content by reconstructing a phenetic tree that provides clusters that can be associated with known phenotypes [90]. Figure 3.1 illustrates the heatmap and hierarchical clustering performed with Ray Surveyor on our dataset. The hierarchical clustering, shown on the left and top axes of the heatmap, defines seven clusters with their colour ranging from green to blue. On the same axes the associated BMI (top axis) and country (left axis) of the individuals are identified by triangles. With regard to the BMI, the most homogeneous clusters were found to be associated with the normal condition in the colours : green (100% normal), red (96.94% normal), magenta (96.77% normal) and yellow (90.54% normal). The two clusters with the most obese individuals are the black and blue clusters with 38.6% and 35.23% respectively of obese individuals. The cyan cluster was represented at 82.22% within the normal cohort. For the country of the samples, the first three clusters (green, red and cyan) are ubiquitously composed of people with Western (Europe and USA) origin with 353 individuals from the Netherlands, 15 from Spain, 17 from Denmark and 12 from the United States (USA). The magenta and yellow clusters are mostly composed of individuals with African origin with 75 from Madagascar (71.4%) and Peruvian origin (9.5%) with the rest of the samples originating from Denmark (17), the Netherlands (2) and the USA (1). Finally, the black and blue clusters are mixed, although predominantly European, with 77 individuals from Denmark, 24 from Spain, 28 from the Netherlands, 9 from Madagascar, 6 from the USA and 1 from Peru.

Selecting the hierarchical clustering to the end of the tree ramification would have yielded additional clusters and certainly improved the overall homogeneity of the clustering. However, there is an interesting demarcation in the broadest clusters between the lower-left group 1 (green, red and cyan colours) and the upper right sections identified as group 2 (magenta, yellow, black and blue colours). Indeed, the clustering suggests the existence of a relationship between the BMI and the clustering of k -mers metagenomics content, even though the complexity that underpins obese conditions maybe too complex to be explained by the gut microbiota alone. Nonetheless, the metagenomics features that distinguish group 1 and group 2 overlap substantially with the one that discriminates normal from obese individuals as identified by the statistical and machine learning analyses of protein annotations. The next sections will present these characteristics, based on taxonomy, protein functions and pathway abundances, that contribute to the delineation of the metagenomes in the context of body fatness.

3.5.2 Taxonomic analyses of the gut metagenomes

Taxonomic analyses are one of the foundations of metagenome investigations using primarily 16S rRNA and WGS experiments. Investigating different taxonomic ranks often yields different results and is greatly influenced by the quality of the reference database, a common occurrence in most genomics studies [347]. We used the MGnify protein database because of its focus on the gut microbiome. It has been recently updated with the numerous protein annotations including Gene Ontology (GO), KEGG pathways, and Clusters of Orthologous Groups (COG) among others [265].

A multitude of microbes found in the gut have not been cultured and characterized [432], thus curation projects such as MGnify are needed to account for those uncharacterized bacterial species available through public WGS metagenomics studies. Understandably not every species is well characterized and therefore only a few proteins have a taxonomic identification down to the species or subspecies. In fact in MGnify, only 10% of the proteins are associated with a taxon down to the genus, and respectively 52%, 60%, 89% and 94% at the family, order, class and phylum taxonomic rank. The taxonomic classification can also be impacted by the protein sequences clustering at 90% identity that can regroup protein homologs found in different species. Nevertheless, the results of the annotations provide a representative portrait of the gut microbiome proteome which holds the essential information of its functional capabilities. Figure 3.2 shows the Bacteria domain cladogram down to the taxonomic family rank. It also indicates the most dominant phyla and families found in the gut metagenomes of 640 healthy and obese individuals. The table 3.1 along with figure 3.2 highlights the different bacterial families with their overall abundances and expression status in obese.

We noted an absence of significant differences when comparing the normal group as defined by the CDC (BMI between 18.5 and 25) with the overweight group (BMI between 25 and 30). Furthermore, including the overweight in the normal cohort yielded very similar p-values when comparing only the normal group with the obese 1,2 and 3 groups ($\text{BMI} \geq 30$) without including the overweight group. This observation remained valid for the other taxonomic ranks and protein annotations analyzed, thus justifying the BMI threshold at 30 to separate the groups in our experiments.

The first taxonomic analysis was at the phylum rank, as many but not all gut metagenomics studies have reported, the significance of their abundance and ratio in explaining obesity [228; 229; 394; 392; 349]. In figure 3.3, we reported the four most abundant phyla, ordered by the normal cohort individual mean, based upon their differential abundance between obese and normal individuals. Each boxplot represents a BMI group as defined by the CDC, except for obesity 3 and 4 that were combined due to the limited number of samples in those two groups. Our results for the *Firmicutes* and *Bacteroidetes* ratio are in agreement with Schwartz *et al.* and contrast to observations reported in Ley *et al.* and Turnbaugh and collaborators. Indeed,

our analyses suggest an unbalanced ratio in favour of the *Bacteroidetes* in obese individuals. Specifically, the mean relative abundance of *Bacteroidetes* was higher by 16% (41% to 25%, $P < .001$) in obese and in contrast *Firmicutes* were lower by 11% (52% to 63%, $P < .001$). On average, those two phyla represented 88.88% of the total microbes sequenced in the gut microbiomes. The third most abundant phylum, *Actinobacteria*, was also significantly ($P < .001$) reduced in obese individuals by 3.8% (from 5.5% to 1.7% on average). *Fusobacteria* were significantly more abundant in obese individuals, despite a low global abundance, going from 0.150% to 0.107% in healthy individuals ($P < .001$). Those four phyla abundance observations were also replicated when investigating the difference between group 1 (mostly normal) and group 2 (predisposition to obesity) from the hierarchical clustering. There were also additional abundant phyla with no significant differences for the BMI but that were capable of differentiating group 1 and group 2 clusters. Indeed, the phyla *Proteobacteria* (5.3x, $P < .001$), *Euryarchaeota* (1.5x, $P < .001$) and *Spirochaetes* (4.7x, $P < .001$) were all relatively enriched in group 2 compared to group 1, indicating potential dysbiosis at a greater propensity in group 2 as generally lower abundant taxa bloom significantly. To appreciate the overall abundances, the cladogram on figure 3.2 is coloured based on the phyla and the important families are highlighted when differently and significantly represented in the metagenomes of normal and obese. The feature extraction from the classification analyses yields an appreciation of which feature (herein the phyla) possesses the most weight in classifying the two cohorts. For the phylum classification, the most important feature is the *Actinobacteria* that outweighs the other phyla by at least twofold in the best performing algorithms. Another important feature that emerged from the classification was the *Deferribacteres* with an overexpression in obese individuals ($P < .001$), although it had a very low mean abundance of 0.0078% in all samples.

The investigation of the microbiome phylogenetic composition at the family taxonomic rank allowed us to further evaluate the important taxa related to body fatness. As in the phylum analysis, we selected the most abundant families that were significantly different in terms of abundance in obese and normal individuals. Figure 3.4 displays the top 9 families ordered by their abundance as found in the healthy group. The first two families *Lachnospiraceae* and *Ruminococcaceae* are from the *Firmicutes* phylum and constitute on average 25.4% and 22.6% of a normal microbiome composition, based on our criteria. Both were underrepresented in obese individuals by 6.7% and 8% respectively on average ($P < .001$). The next two most prevalent *Firmicutes* families were the *Clostridiaceae* and the *Eubacteriaceae* (both at $\approx 7\%$) and did not show significant differences between the obese and normal conditions yet had a significant difference between the group 1 and group 2 clusters of figure 3.1. The most abundant *Bacteroidetes* family was *Bacteroidaceae* with an overall abundance of 21.8% (second overall after the *Lachnospiraceae*) and a greater expression in obese gut metagenomes by more than 15% on average (19.9% \rightarrow 35.2%, $P < .001$). The next most abundant families from the *Bacteroidetes* phylum was *Porphyromonadaceae* with an overall abundance of 4.5% and overexpressed in obese (4.1% \rightarrow 6.9%, $P < .001$), *Rikenellaceae* at 1.7% with no significant

difference, the *Prevotellaceae* at 0.31% and the *Flavobacteriaceae* at 0.25%. Interestingly, the *Prevotellaceae* is the only reported *Bacteroidetes* in the top 20 families that is underexpressed in obese metagenomes (0.32% \rightarrow 0.27%, $P < .001$). From the classification analyses, the *Ruminococcaceae* and *Micrococcaceae* are two families overrepresented in the predictions for obesity, and that result was obtained with multiple algorithms.

We also analyzed the most abundant genera, even though only 10% of the proteins in the MGnify database are annotated down to the genus rank, thus one needs to be cautious in overinterpreting the results. The six most abundant and differentially expressed genera, still ordered by their abundance in healthy individuals, are shown in supplementary figure 1 (figure A9). The three most abundant are *Blautia*, *Dorea* and *Lachnospiraceae*, which all come from the *Lachnospiraceae* family and constitute respectively on average 39.6%, 17.5% and 15.1% of the overall abundance of the gut metagenomes in our experimental group. The difference in expression of the *Blautia* and *Dorea* genera is in line with their family (*Lachnospiraceae*) tendency as they were underexpressed in obese individuals, respectively by 13.6% (41.4% \rightarrow 27.8%, $P < .001$) and by 3.6% (17.9% \rightarrow 14.8%, $P < .001$). In contrast, the *Lachnospiraceae* and *Butyrivibrio* genera from the same family were overexpressed in obese, respectively by a mean difference of 4.1% and 6.8% ($P < .001$). The next genus with the most identified proteins and being significantly different is *Bacillus* from the *Bacillaceae* family and similarly is overexpressed in obese by a mean margin of 3.4% ($P < .001$). Finally, the last of the six genera of figure supp. 1 (figure A9) is *Alloprevotella*, an overexpressed *Prevotellaceae* even though the family was on average underexpressed in obese individuals ($P < .001$). The top five genera identified in the machine learning analyses are conserved in the two best performing algorithms (LightGBM, Random Forest), and is constituted of *Chryseobacterium*, *Blautia*, *Butyrivibrio*, *Flavobacterium* and *Oceanicola* genera.

3.5.3 Functional analyses

The functional analyses is comprised of the abundances of protein annotation features, such as their Enzyme Commission (EC) numbers, their Gene Ontology (GO) characterization, their Clusters of Orthologous Group (COG) association and their associated pathway and module from the KEGG database.

The results for the KEGG pathways report a multitude of significant pathways (with $P < .001$) that are differentially expressed in the obese and normal cohorts. Figure 3.5 shows the top twenty based on the lowest p-values as reported by the statistical tests. Interestingly, the most significant pathway is the butyrate metabolism which is decreased in obese individuals ($P < .001$) and been shown previously to impact energy homeostasis in mice when comparing germ-free and normal mice [97]. Indeed, butanoate or butyrate is a short-chain fatty acid (SCFA) that has been considered as one of the main products of fermentation by bacteria in the colon [226]. Butyrate can regulate gene expression and could serve as a preferential

energy source for the colonic epithelial cells (colonocytes) [131; 330; 331; 32; 47; 77]. SCFA are generally considered the result of resistant starch and dietary fiber fermentation from bacteria in the gut and are used in glucose and lipid biosynthesis [418; 419; 166].

The second most important pathway from the KEGG database analysis is the microbial metabolism in diverse environmental pathways (ko01120) which are also found to be decreased in the obese cohort. These pathways are related to several others that are involved in the metabolism or degradation of different metabolites, and which are also significantly underexpressed in obesity. To name a few, the glycolysis and gluconeogenesis pathway (ko00010), the glycine, serine and threonine metabolism pathway (ko00260), the carbon fixation pathway in prokaryotes (ko00720), and the nitrogen metabolism pathway, are all related to microbial metabolism and indeed underexpressed in the microbiome associated with obesity. This is a potential indication of metabolism efficiency reduction in obese individuals taking root in the gut microbes.

The importance of the glycolysis and gluconeogenesis pathway in obesity could be linked, one to food intake and satiety, and second as a key factor in insulin sensitivity [266]. Enzymes such as glucokinase (EC : 2.7.1.2) and pyruvate kinase (EC : 2.7.1.40), and their associated GO activities (GO :0004340, GO :0004743) were also found to be decreased in obese metagenomes. The glycine, serine and threonine metabolism pathway has also been studied with relationship to obesity. First, it has been suggested that the gut microbiota plays an important role in glycine availability for the host which would be impacted by diet and essential to multiple metabolic pathways of the host [9]. Second, the three amino acids were also significantly more available for the host in germ-free mice compared to wild type, suggesting the implication of the gut microbiota in their metabolism and availability [253]. Finally, threonine as well as cysteine has been shown to improve protection against colitis in rats by promoting mucin secretion and which is very probably linked to beneficial bacteria in the gut microbiota [371]. The carbon metabolism (ko01200) and fixation pathways (ko00720) are both underexpressed in obese individuals and the annotated proteins mainly come from the *Ruminococcaceae* family, a *Firmicutes* also less abundant in the obese. It has previously been demonstrated that not all bacterial species can utilise the same carbon source to produce fermented products beneficial to the host, such as SCFA [248]. Another noteworthy pathway that is less abundant in obese individuals is the one involved in the biosynthesis of amino acids (ko01230), which is also related to the metabolism of SCFA by providing the required amino acids for their synthesis [26; 319; 274].

On the other hand, the most important pathway for obesity delineation from the classification analyses was the selenocompound metabolism pathway (ko00450) as it appears as the top discriminant feature with different algorithms. It has been suggested by Hrdina and collaborators, when experimenting on mice, that bacteria compete with the host for selenium when availability becomes limiting [170]. A decreased expression in obese individuals could

indicate a selenium-deficient diet as fewer bacteria involved in its metabolism were able to thrive. The gene ontology results also show evidence of selenocompound metabolism in relationship with obesity. Indeed, several activities involving selenocompounds were significantly less abundant in obese individuals such as the L-seryl-tRNA^{Sec} selenium transferase activity (GO :0004125), the selenocysteinyI-tRNA^(Sec) biosynthetic process (GO :0097056), the transferring selenium-containing groups transferase activity (GO :0016785), the selenate reductase activity (GO :0033797), and the selenocysteine incorporation, insertion sequence binding, metabolic process and biosynthetic process (GO :0001514, GO :0035368, GO :0016259, GO :0016260) ($P < .001$).

The vast majority of the top significant pathways were underexpressed in the obese cohort with the notable exceptions of the peroxisome pathway (ko04146), the NOD-like receptor signalling pathway (ko04621), the ferroptosis pathway (ko04216) and the biofilm formation in *Vibrio cholerae* pathway (ko05111). There is evidence that suggests that these four pathways could be related to immune response and pro-inflammatory reactions [93; 342; 313; 439]. Interestingly, the peroxisome pathway which is also involved in lipid metabolism has been suggested as an important factor to maintain gut epithelium homeostasis and renewal in *Drosophila* [94]. Indeed, dysfunctional peroxisomes in the host gut epithelial cells would trigger Tor kinase-dependent autophagy that would increase cell death and promote instability at the host-microbe interface in the gut. Similarly, the ferroptosis pathway would also be involved in intestinal epithelial cell death and could even lead to ulcerative colitis [428]. Even though those two pathways are usually considered as part of host cells, we found several bacterial proteins in the MGnify database that were classified as part of these pathways. The main enzyme that was predominantly part of both pathways was the long-chain-fatty-acid-CoA ligase (EC : 6.2.1.3), which also showed an increased abundance in obese individuals and could explain the pathway results. The NOD-like receptor signalling pathway genes characterized in the MGnify database are for the most part related to the thioredoxin system (such as *trxA*) and are predominantly found in the *Bacteroidaceae* and *Porphyromonadaceae* families, both overexpressed in obese gut metagenomes. Different studies have investigated the role of tetrathionate in the gut microbiota, including one suggesting that an upregulation of the thioredoxin reductase *trxA* may modulate the gut microbiome during inflammation by regulating the levels of tetrathionate [270], and another one from Winter et al. showing a competing growth advantage in inflamed gut for *Enterobacteriaceae*, such as *Salmonella* in presence of tetrathionate [417].

The COG results are also in agreement with the other annotation results, as the main significant categories are related to energy production and conversion and the amino acid and nucleotide transport and metabolism, and are all underrepresented in obese metagenomes ($P < .001$). However, the COG annotations have a very broad categorization and do not accommodate for more specific functional analyses.

Finally, KEGG modules are another annotation resource that provides a functional annotation

unit for the proteins, although they are related to KEGG pathways. The most significant module reported in the statistical test is associated with purine degradation (M00546) and is less abundant in obese individuals. The module is part of the purine metabolism KEGG pathway, which has not proved to be different in the pathway analyses. Interestingly, the degradation of purine is associated with gout, a condition that is characterized by the accumulation and reduced excretion of uric acid [152]. Indeed, Guo and collaborators found a disorder in purine degradation and butyric acid biosynthesis in the gut metagenomes. In agreement with our study, the butyric acid biosynthesis was less abundant in gout. However, the purine degradation pathway was enriched in gout and depleted in obese individuals. Nevertheless, body fatness (visceral fat) has been reported to strongly correlate with gout [224]. In the context of obesity, the lack of bacterial species that metabolize purine into uric acid may contribute to the condition. Several other modules involved in the transport of metabolites such as cystine (M00234), glutamine (M00228), rhamnose (M00220), molybdate (M00189), glutamate (M00233) and others were also significantly depleted in obese individuals. The best performing classification algorithm also reported the chorismate reaction of the shikimate pathway as the most significant module that was enriched in normal individuals. The shikimate pathway is found exclusively in microorganisms and plants and is mainly dedicated to the production of aromatic amino acids (phenylalanine, tyrosine, and tryptophan) in bacteria [163]. Evidence has also shown that perturbation in the shikimate pathway is diet-related [293].

3.5.4 Obese metagenome classification

We also sought to classify the two groups : the obese, with a BMI of ≥ 30 , versus normal and overweight combined with a BMI < 30 . The same protein annotation features that were used in the statistical analyses were used to train and evaluate different machine learning algorithms : xgboost, lightgbm, random forest, svm, decision tree, adaboost and the SCM [59; 200; 168; 158; 41; 161]. We hoped that the MGnify database, a very recently updated protein annotation database, which includes information from several other important databases, like KEGG, EC, and GO [191; 21; 18], would enable us to improve the annotation coverage of the metagenomic proteomes.

Similar work had been conducted in Pasolli and collaborators, for the meta-analysis of several diseases associated with the gut microbiota. Their results for the obesity condition were solely based on the taxonomic profiling obtained from Metaphlan2 [390] and the discrimination was made between lean (BMI ≤ 25) and obese (BMI > 30) individuals without including the overweight condition. Unfortunately, obesity [223] was the least predictable condition as several others such as liver cirrhosis [317], colorectal cancer [436], inflammatory bowel diseases (IBD) [315], and type 2 diabetes [316; 194] reported superior classification scores. Nevertheless, based on the premise that cross-study sample inclusion was an improving factor, we built the obese and non-obese cohorts and evaluated classification accuracy. We did improve the overall results

but not by a large margin for the cirrhosis, colorectal cancer and IBD conditions. However we obtained significant improvement for type 2 diabetes results and the obesity classification. The evaluation results for the machine learning algorithm are reported in table 3.2. The tables are separated by protein annotation features, either functional or taxonomic, as statistically analyzed in the previous sections. The best performing features, based on the F1 score results, are the COG annotation (0.66) closely followed by the taxonomic genus (0.65), the gene ontology (0.64), the enzyme commission (0.63) and the KEGG module (0.62) annotations. From all the evaluated algorithms, lightgbm (gradient boosting) and random forest were the ones that yielded the best results overall. All individual features achieved better scores than the one reported in Pasolli et collaborators with the best F1 score being at 0.55 for the support vector machine. Our improved results could be attributed to the inclusion of samples from several studies, especially for the normal cohort, and the usage of an updated metagenomics database for the protein annotations. The optimized results allowed the extraction of important features in the discrimination between obese and normal gut metagenomes as reported in the functional analyses section.

3.6 Discussion

In this study, we investigated the gut metagenome of 640 normal, overweight and obese individuals based on their BMI. We evaluated their metaproteomes seeking the difference in the functional and taxonomic annotation that would be additional indicators of the obesity condition. Overall, we found that the overweight group (BMI between 25 and 30) was not significantly different from the normal group (BMI between 18.5 and 25). Indeed, differences were readily noticed when comparing the obese group with the non-obese group, that is with a BMI cutoff at 30. It is important to take into account that BMI measurements alone have some limitations as a reflection of the percentage of body fat since its association varies between different ethnic groups, such as the Asian population when compared to Caucasian. [141]. Nonetheless, BMI is still used by the World Health Organisation (WHO) for the evaluation of obesity and excess weight. We used it for this study due to its wide availability and acceptability. The k -mer analysis showed two broad groups with a greater concentration of obese individuals in the second group. However, the clusters were also greatly representative of the ethnicity of the individuals with a distinction between Western and non-Western origins. It suggests that factors related to ethnicities such as diet, lifestyle and geography are contributing to shaping the gut microbiome, as previously suggested [154; 133]. No clusters were predominantly represented by obese metagenomes, with potential explanations being the unbalanced nature of our datasets, extrinsic factors such as diet, genetics and lifestyle and importantly that the BMI is not perfectly representative of body fatness, hence our effort to find better biomarkers [282].

The overall relative abundances of the major phyla, such as the *Firmicutes*, *Bacteroidetes*

and *Actinobacteria*, were in agreement with previous studies. With regard to their relative abundances in normal versus obese individuals, contrasting results have been reported in the literature [228; 229; 394; 392; 349; 223]. Our results add support to an enriched proportion of *Bacteroidetes* in obese individuals when comparing the *Bacteroidetes* to *Firmicutes* ratio. Moreover, it is the *Actinobacteria* that best discriminated the obese and the normal groups, with a depleted abundance in the obesity group. *Actinobacteria* has been reported to be represented mainly by the *Bifidobacterium* genus in the gut metagenome [17] and their decrease is associated with several conditions in addition to obesity, such as types I and II diabetes, cystic fibrosis, hepatitis B and *Clostridium difficile* infections [15]. At the family taxonomic rank, the *Lachnospiraceae* (24.6%) and *Ruminococcaceae* (21.6%) were the major representatives of the *Firmicutes* phylum whereas the *Bacteroidaceae* family (21.8%) was the most abundant *Bacteroidetes*. Together, those three families represented an overall abundance of 68% of the metagenomes on average and their relative abundances were in line with their respective phyla when comparing the obese and non-obese groups.

The functional proteome analyses revealed the importance of short-chain fatty acid (SCFA) metabolism in the healthy condition in comparison with obesity. Indeed, butyrate metabolism was the most statistically significant pathway depleted in obese individuals. Butyrate along with propionate and acetate are SFCA produced by the fermentation of dietary fibers and starch by gut bacteria and play an important role in energy availability for the epithelial cells in the colon [236]. Other pathways, such as amino acid biosynthesis, were also decreased in obese metagenomes and are known to be related to SCFA production by providing the necessary building blocks of their synthesis. Another discriminant example is the bacterial shikimate pathway, which essentially produces aromatic amino acids, and from which one reaction (from KEGG modules) proved to be the most important feature in the machine learning analyses to separate the obese from the normal cohort. Furthermore, we also reported the importance of the selenocompound metabolism pathway, which is also related to diet and characterized by a selenium-deficient food intake from individuals with an obesity condition. On the other hand, the pathways that were more abundant in obese individuals were mostly involved in immune response and pro-inflammatory reactions.

Overall, by analyzing metagenomes collected from different studies, we were able to identify significant changes in obese versus non-obese individuals. By using protein annotations, we drew the taxonomic portrait and functional capabilities of the metagenomes that assist in interpreting the impact of the obesity condition on the gut microbiota. Additional metagenomics experiments on obesity with diverse cohorts of individuals, controlled diets and other measurements of body fatness would enhance our comprehension of the complex interactions between human obesity and the gut microbiome.

Figures / Tables

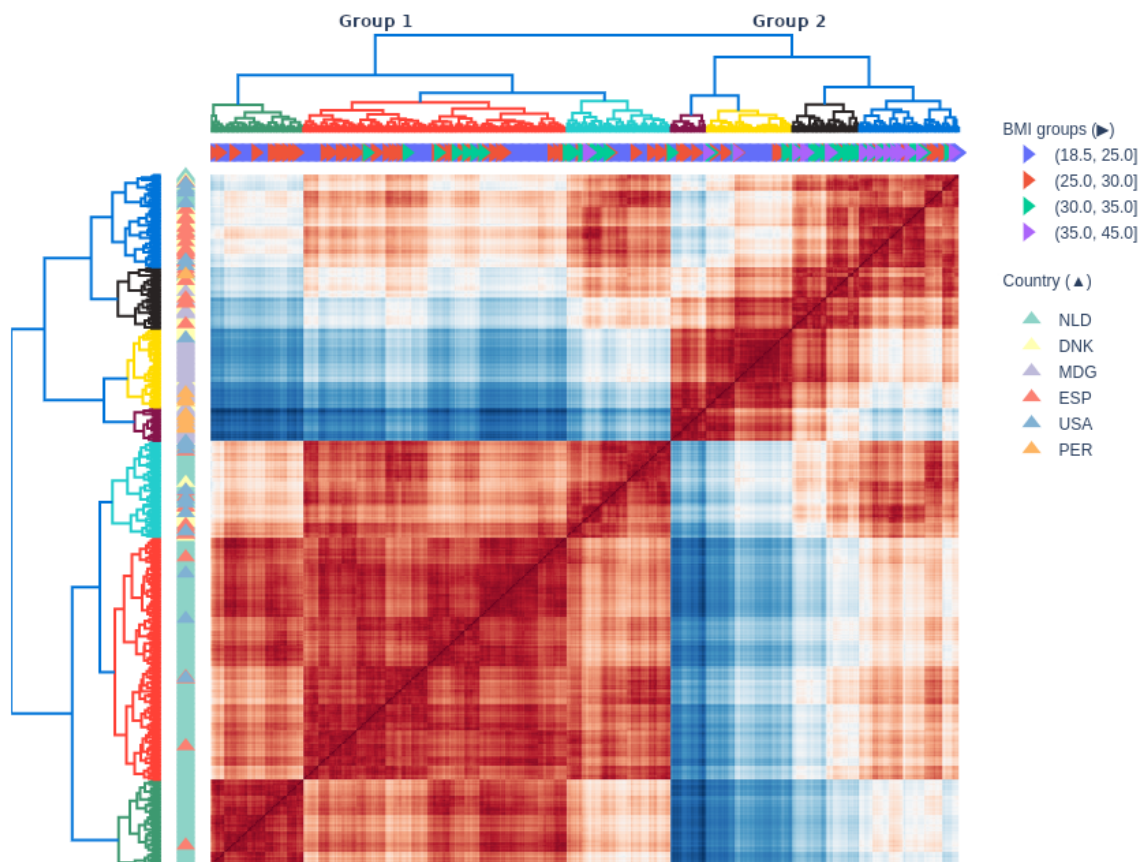


FIGURE 3.1 – Heatmap and hierarchical clustering of the metagenome based on their shared DNA content (k -mers of length 31). Red color indicate close similarity between metagenome as blue color indicate more dissimilarity. BMI groups are indicated by a right triangle in the upper portion of the figure and country of the individuals on the left side by upper triangles.

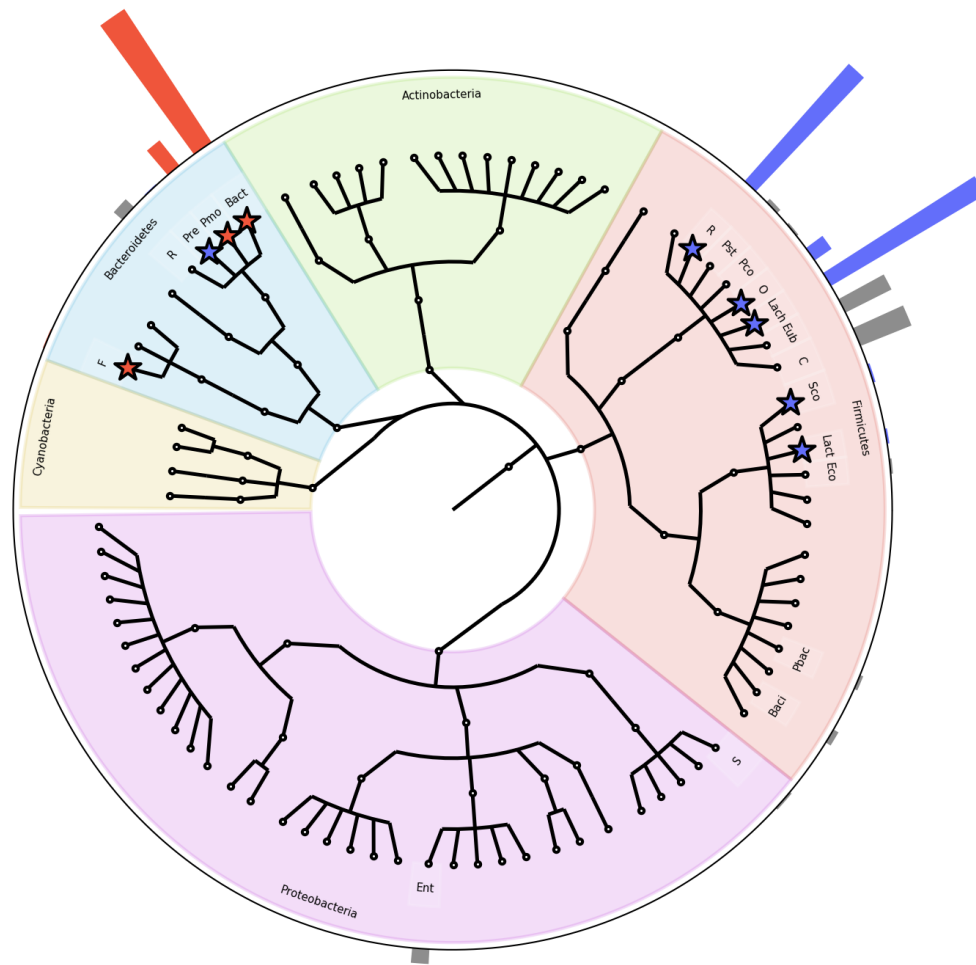


FIGURE 3.2 – Bacteria cladogram of the gut metagenomics data found in 640 individuals down at the family taxonomic rank. Red bars indicate higher relative abundance of the taxa in obese individuals, while blue bars indicate higher representation in normal individuals. Stars indicate that the relative abundance was significantly different in both cohorts.

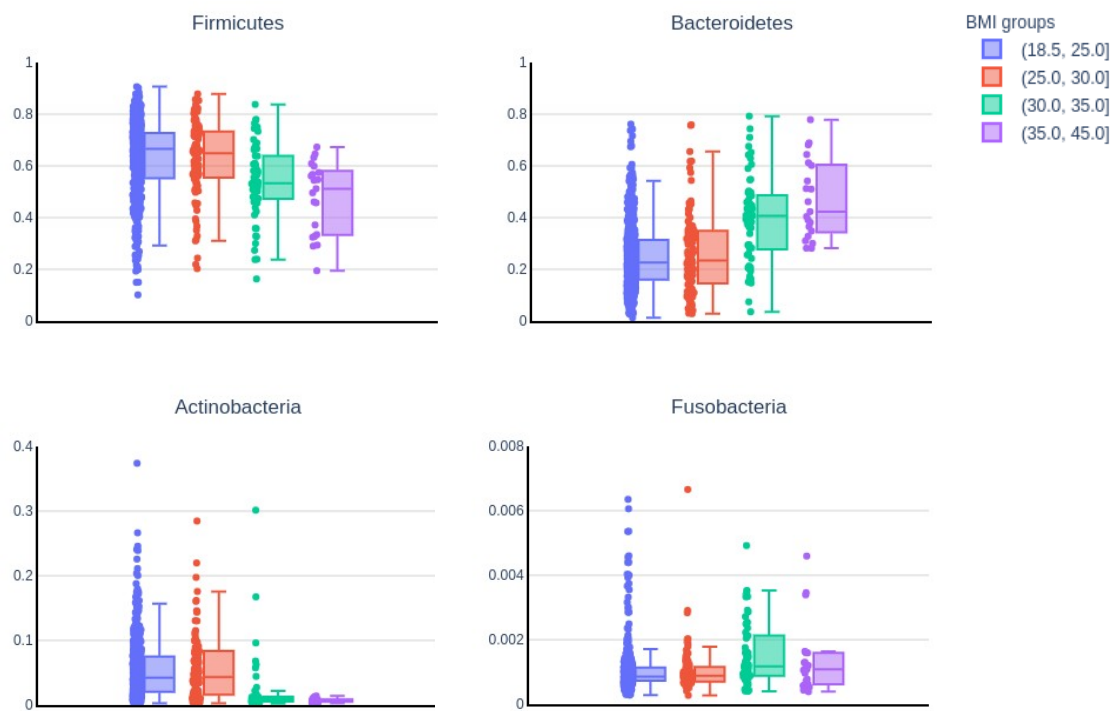


FIGURE 3.3 – The top 4 most abundant phyla with significant changes in obese and non-obese gut microbiota.

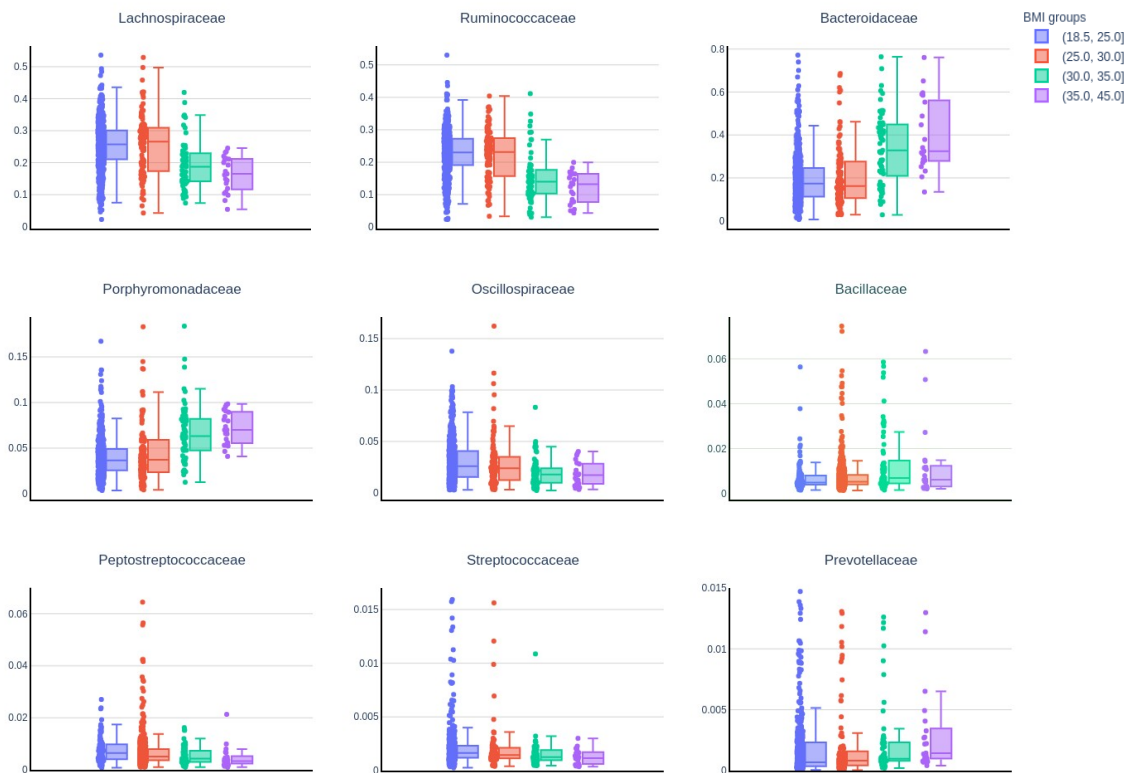


FIGURE 3.4 – The top 9 most abundant families with significant changes in obese and non-obese gut microbiota.

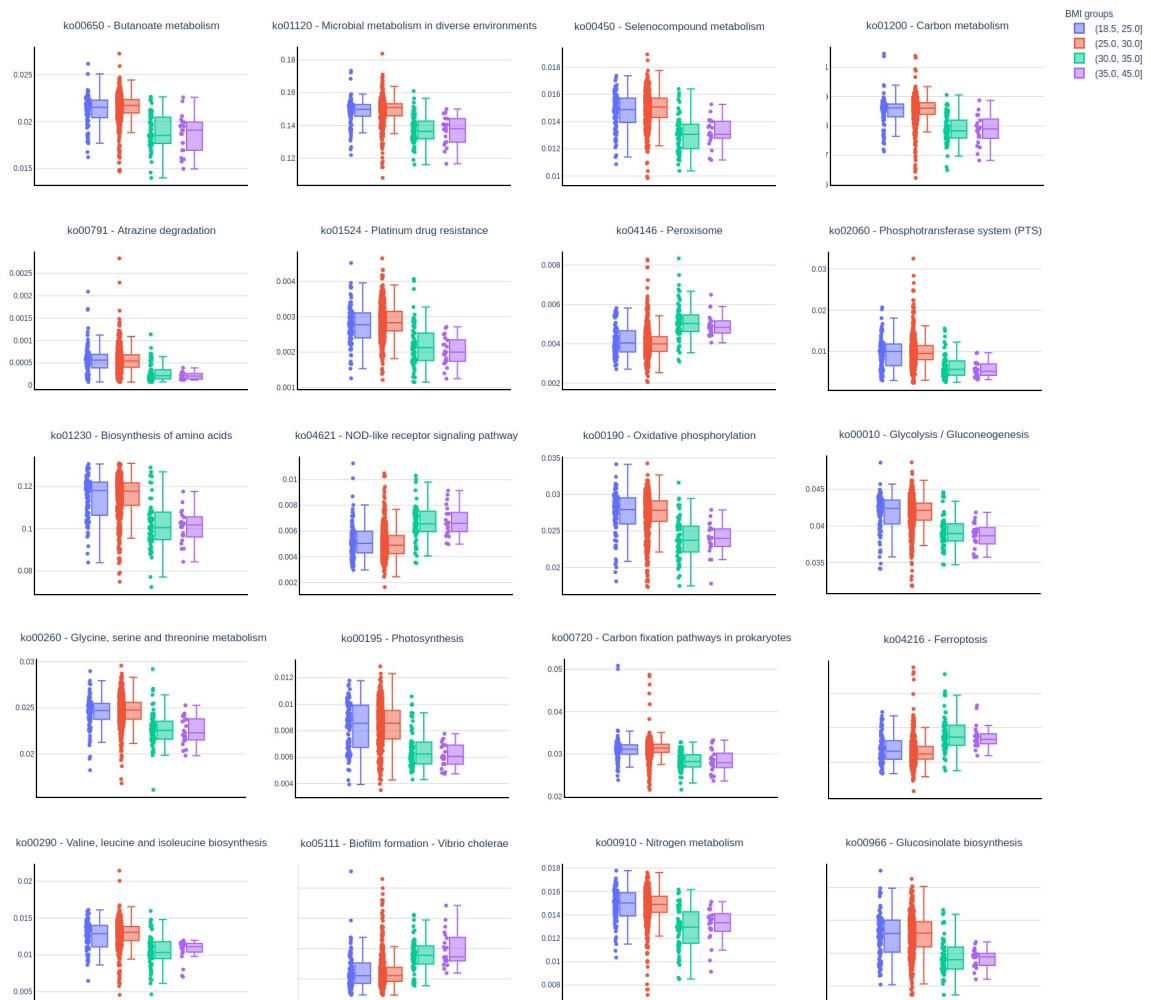


FIGURE 3.5 – The top 20 KEGG pathways which show the most significant changes in abundances in obese and non-obese gut microbiota.

TABLEAU 3.1 – Top 20 families based on their overall abundance in our dataset. The family codes serve as a reference for the families displayed in figure 3.2.

Family Code	Family	Phylum	Overall Abundance (%)	Obese Expression (↑/↓)
Lach	Lachnospiraceae	Firmicutes	24.6	↓
Bact	Bacteroidaceae	Bacteroidetes	21.8	↑
R	Ruminococcaceae	Firmicutes	21.6	↓
C	Clostridiaceae	Firmicutes	7.6	-
Eub	Eubacteriaceae	Firmicutes	7.1	-
Pmo	Porphyromonadaceae	Bacteroidetes	4.5	↑
O	Oscillospiraceae	Firmicutes	2.9	↓
Ent	Enterobacteriaceae	Proteobacteria	2.0	-
Rik	Rikenellaceae	Bacteroidetes	1.7	-
Baci	Bacillaceae	Firmicutes	0.85	↑
Pst	Peptostreptococcaceae	Firmicutes	0.74	↓
Pbac	Paenibacillaceae	Firmicutes	0.62	-
S	Sutterellaceae	Proteobacteria	0.44	-
Lact	Lactobacillaceae	Firmicutes	0.40	-
Pco	Peptococcaceae	Firmicutes	0.36	-
Eco	Enterococcaceae	Firmicutes	0.32	-
Sco	Streptococcaceae	Firmicutes	0.31	↓
Pre	Prevotellaceae	Bacteroidetes	0.31	↓
F	Flavobacteriaceae	Bacteroidetes	0.25	↑
Sip	Siphoviridae	(Virus superkingdom)	0.23	↓

TABLEAU 3.2 – Machine learning obesity classification results from metagenomics’ protein annotations

Taxa Phylum							
score/algo	xgboost	lightgbm	random forest	support vector machine	decision tree	adaboost	set covering machine
test_balanced_accuracy	0.706	0.789	0.805	0.500	0.834	0.712	0.630
test_f1	0.498	0.561	0.544	0.049	0.581	0.493	0.378
test_roc_auc	0.849	0.847	0.855	0.768	0.840	0.754	0.630

Taxa Family							
score/algo	xgboost	lightgbm	random forest	support vector machine	decision tree	adaboost	set covering machine
test_balanced_accuracy	0.775	0.766	0.842	0.799	0.775	0.696	0.705
test_f1	0.569	0.602	0.591	0.511	0.498	0.511	0.477
test_roc_auc	0.868	0.887	0.880	0.845	0.756	0.840	0.705

Taxa Genus							
score/algo	xgboost	lightgbm	random forest	support vector machine	decision tree	adaboost	set covering machine
test_balanced_accuracy	0.754	0.805	0.726	0.500	0.723	0.683	0.738
test_f1	0.576	0.640	0.534	0.049	0.428	0.464	0.519
test_roc_auc	0.841	0.837	0.871	0.815	0.766	0.810	0.738

Enzyme commission

score/algo	xgboost	lightgbm	random forest	support vector machine	decision tree	adaboost	set covering machine
test_balanced_accuracy	0.788	0.804	0.771	0.500	0.776	0.759	0.648
test_f1	0.630	0.632	0.608	0.177	0.489	0.580	0.402
test_roc_auc	0.881	0.871	0.889	0.870	0.775	0.844	0.648

Gene Ontology

score/algo	xgboost	lightgbm	random forest	support vector machine	decision tree	adaboost	set covering machine
test_balanced_accuracy	0.745	0.800	0.702	0.500	0.766	0.692	0.609
test_f1	0.582	0.652	0.518	0.049	0.523	0.478	0.319
test_roc_auc	0.882	0.873	0.894	0.851	0.710	0.810	0.609

COG

score/algo	xgboost	lightgbm	random forest	support vector machine	decision tree	adaboost	set covering machine
test_balanced_accuracy	0.824	0.813	0.830	0.669	0.813	0.809	0.806
test_f1	0.604	0.615	0.665	0.412	0.544	0.610	0.627
test_roc_auc	0.890	0.886	0.897	0.681	0.810	0.813	0.806

KEGG pathway

score/algo	xgboost	lightgbm	random forest	support vector machine	decision tree	adaboost	set covering machine
test_balanced_accuracy	0.769	0.763	0.766	0.500	0.785	0.730	0.708
test_f1	0.587	0.587	0.603	0.049	0.514	0.554	0.486
test_roc_auc	0.860	0.869	0.889	0.837	0.791	0.855	0.708

KEGG module

score/algo	xgboost	lightgbm	random forest	support vector machine	decision tree	adaboost	set covering machine
test_balanced_accuracy	0.768	0.835	0.687	0.500	0.790	0.737	0.718
test_f1	0.583	0.621	0.499	0.049	0.528	0.553	0.505
test_roc_auc	0.860	0.895	0.893	0.872	0.798	0.879	0.718

Conclusion et perspectives

L'objectif principal de ce doctorat était de contribuer au développement de méthodes d'analyses en génomique et métagénomique avec un intérêt particulier porté à la comparaison de génomes bactériens dans un contexte d'élucidation phénotypique et fonctionnelle.

La première hypothèse de recherche a mené à l'implémentation du logiciel Ray Surveyor pour permettre la comparaison *de novo* de génomes et métagénomes bactériens. Ray Surveyor s'est inséré dans la suite logiciel Ray pour bénéficier de ses avancées techniques en matière de calcul distribué [33; 34]. L'étude sur Ray Surveyor a paru dans le journal *Molecular Biology and Evolution* et présente plusieurs analyses génomiques et phénotypiques à grande échelle sur des espèces bactériennes. Les analyses incluent entre autres plusieurs éléments génétiques ayant un intérêt clinique, tels que les facteurs de virulences, les gènes de résistance aux antibiotiques et les éléments mobiles pouvant être impliqués dans leur dissémination [90]. La méthode fut aussi validée sur un jeu de données synthétiques en la comparant aux méthodes traditionnelles de comparaison de génomes, tel la phylogénie et la moyenne de pourcentage d'identité. Pour mettre en perspective les travaux sur Ray Surveyor, plusieurs compteurs de k -mers, tels que KMC2, Jellyfish, DSK, pour ne nommer que ceux-là, offrent aussi de très bonnes performances avec une utilisation en ressource informatique très optimisée. Ray Surveyor offre cependant plus d'options que le simple comptage de k -mers avec la création de matrices de similarité et de distance pouvant être utilisées directement avec des méthodes de «clustering» (agrégation). À titre d'exemple, la comparaison directe du contenu génomique dans des métagénomes, en termes de présence et absence de k -mers, s'avère grandement facilité avec le logiciel Ray Surveyor. Dans les pistes à explorer pour l'amélioration de Ray Surveyor, il y aurait l'utilisation des «minimizers» pour le stockage des k -mers qui pourrait réduire la quantité de mémoire nécessaire pour le calcul des matrices de distances ou encore l'utilisation de disque comme espace de stockage temporaire au calcul. Un autre avantage dont Ray Surveyor pourrait bénéficier pour la comparaison de métagénomes serait de faire un calcul de distance qui tiendrait compte du nombre de k -mers présents dans chacun des échantillons et non seulement de leur présence ou absence. Le développement d'un modèle mathématique serait alors nécessaire, puisque les lois s'appliquant à la simple présence ou absence des k -mers ne tiendraient plus avec des données quantitatives dans les comparaisons.

La deuxième hypothèse de recherche a mené à l'implémentation du logiciel kAAmer pour permettre l'identification de séquences de protéines au sein de génomes et métagénomes. L'article du deuxième chapitre présente la méthodologie utilisée dans kAAmer qui consiste brièvement à utiliser des k -mers d'acides aminés pour l'identification de protéines. Un «benchmark» fut d'ailleurs réalisé pour comparer l'efficacité de kAAmer avec d'autres logiciels de pointe en identification de séquences, tels que DIAMOND [43] et BLAST [46]. KAAmer a démontré des performances très intéressantes, particulièrement pour le temps requis d'exécution qui penchait en faveur de celui-ci plus le nombre de requêtes devenait important. Les principales améliorations pratiques pour le logiciel kAAmer seraient le temps de calcul requis pour bâtir une base de données ainsi que l'espace disque nécessaire pour stocker celle-ci. Les nouvelles versions de l'engin clé-valeur utilisé par kAAmer (Badger [92]) supportent maintenant la compression à la volée. Il serait intéressant de refaire les tests de performance avec une nouvelle version de Badger et la fonctionnalité de compression activée pour voir la réduction de l'espace disque utilisé ainsi que l'impact sur la performance de la recherche des clés dans la base de données. D'autres fonctionnalités intéressantes pourraient aussi être implémentées dans kAAmer pour ajouter à l'attrait du logiciel. Par exemple, un «clustering» automatique à partir des k -mers des protéines contenues dans une base de données pourrait apporter une visualisation intéressante de celle-ci en plus de permettre l'investigation des annotations protéiques à partir des regroupements créés. De plus, pouvoir comparer des génomes ou métagénomes entre eux à partir des k -mers en acides aminés, de manière similaire à Ray Surveyor, pourrait aussi être un type d'analyse à explorer.

La troisième et dernière hypothèse de recherche était testé en analysant un ensemble de métagénomes intestinaux d'individus présentant des indices de masses corporelles normaux, en surpoids et obèses. Comme mentionnés précédemment, plusieurs études sur le sujet ont rapporté des résultats parfois contradictoires principalement en lien avec le ratio des abondances relatives des principales phyla *Bacteroidetes* et *Firmicutes*. Nous avons donc appliqué les méthodes développées durant ce doctorat pour réaliser les analyses des métagénomes et explorer les potentielles relations entre le microbiote intestinal et la condition d'obésité. Nos résultats concernant le ratio des *Bacteroidetes* et *Firmicutes* se sont avérés en accord avec l'étude de Schwartz et collaborateurs, soit un enrichissement en *Bacteroidetes* chez les individus obèse [349]. Les résultats se sont aussi avérés significatifs pour une multitude de fonctions et sentiers métaboliques encodés par les métagénomes. Parmi les fonctions dignes de mention, on dénombre celles associées au métabolisme des acides gras à chaîne courte («shot-chain fatty acids» (SCFA)). Il est intéressant de noter que les SCFA ont fait l'objet de beaucoup d'études sur leur implication potentielle dans différentes conditions liées à la santé humaine en plus de l'obésité [419; 166; 51; 205; 349]. D'un autre côté, on observait aussi dans nos analyses une démarcation importante entre les métagénomes des individus provenant de différentes ethnicités. Pour apporter des conclusions plus définitives sur les liens causals entre le métagénome intestinal et la condition d'obésité, des études additionnelles seront requises avec un contrôle

strict sur plusieurs facteurs ayant aussi potentiellement un impact sur la composition du microbiome, tels que l'ethnicité, la diète, l'activité physique et les antécédents génétiques des individus.

Annexes

Figures

FIGURE A1 – Exemple de rapport de résultats de fastp sur une lecture (ERR321632) d'un métagénome du microbiote intestinal humain.

(a) Sommaire des résultats

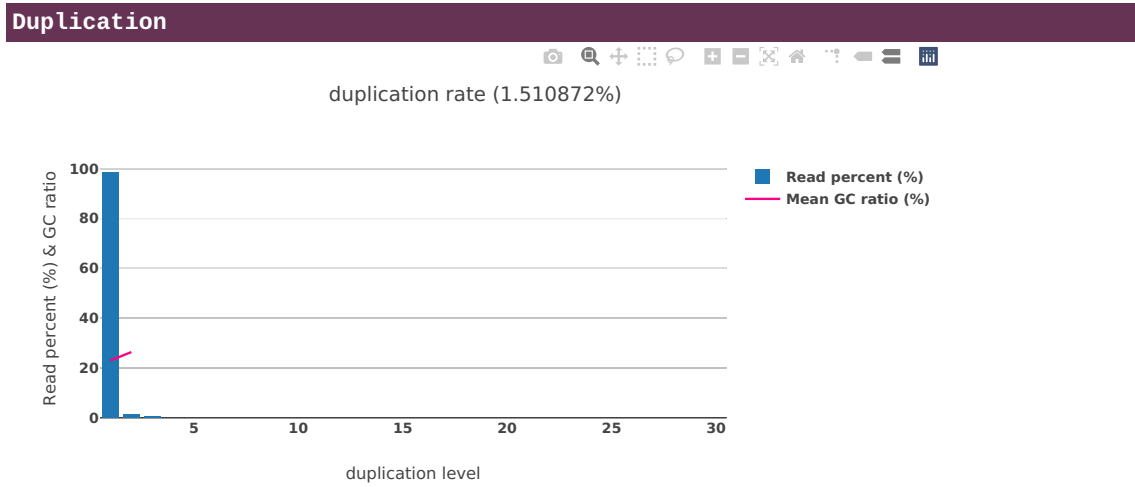
fastp report

Summary	
General	
fastp version:	0.19.5 (https://github.com/OpenGene/fastp)
sequencing:	paired end (90 cycles + 90 cycles)
mean length before filtering:	90bp, 90bp
mean length after filtering:	89bp, 89bp
duplication rate:	1.510872%
Insert size peak:	148
Before filtering	
total reads:	94.429580 M
total bases:	8.498662 G
Q20 bases:	7.652990 G (90.049349%)
Q30 bases:	7.005641 G (82.432276%)
GC content:	46.034940%
After filtering	
total reads:	72.649732 M
total bases:	6.537940 G
Q20 bases:	6.307767 G (96.479427%)
Q30 bases:	5.955604 G (91.092965%)
GC content:	44.145255%
Filtering result	
reads passed filters:	72.649732 M (76.935354%)
reads with low quality:	21.779848 M (23.064646%)
reads with too many N:	0 (0.000000%)
reads too short:	0 (0.000000%)

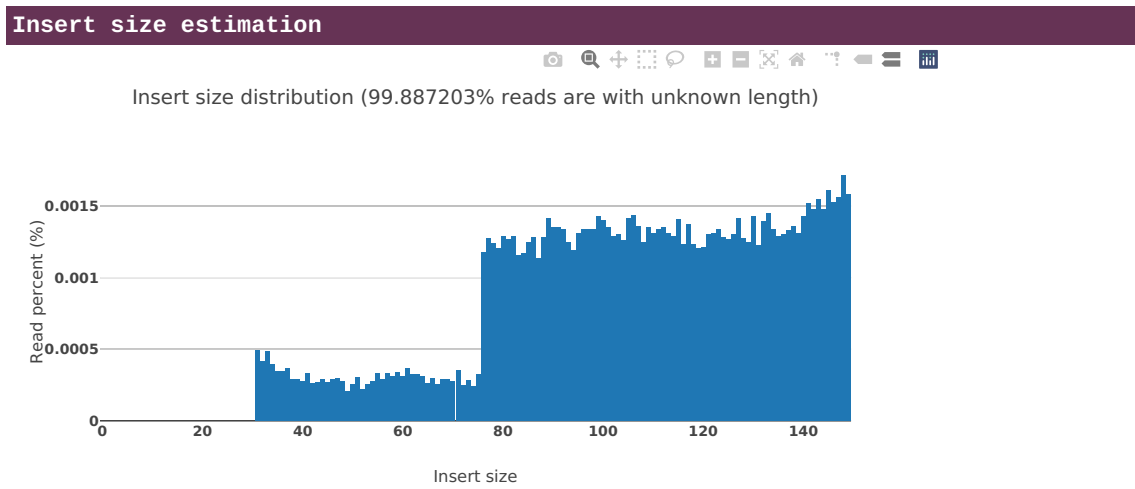
(b) Adapteur ou mauvaise ligature dans la lecture 1

Adapters	
Adapter or bad ligation of read1	
The input has little adapter percentage (~0.006819%), probably it's trimmed before.	
Sequence	Occurrences
A	434
AG	426
AGA	358
AGAT	401
AGATC	413
AGATCG	411
AGATCGG	392
AGATCGGA	391
AGATCGGAA	375
AGATCGGAAG	405
AGATCGGAAGA	348
AGATCGGAAGAG	367
AGATCGGAAGAGC	366
AGATCGGAAGAGCA	348
other adapter sequences	7586

(c) Duplication des lectures



(d) Estimation de longueur des insertions entre les lectures pairées

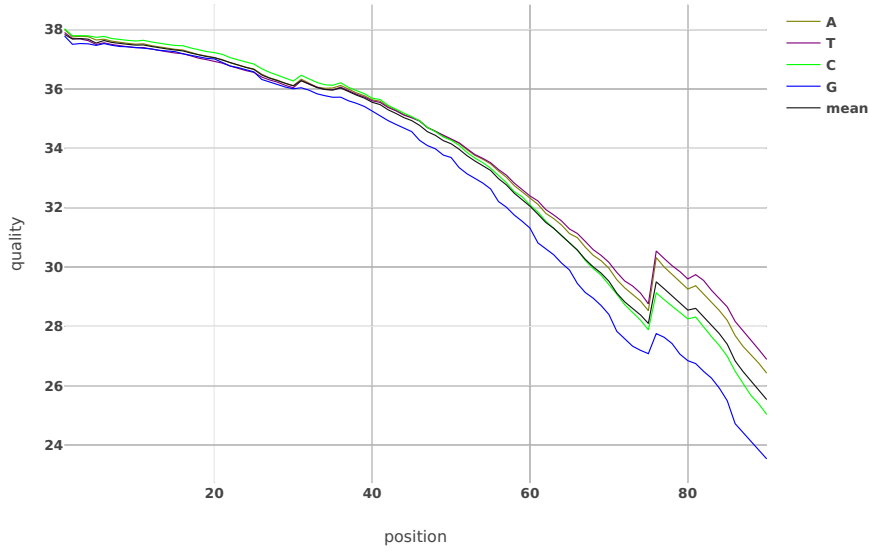


(e) Qualité, contenu en base et compte de k-mers de la lecture 1 avant filtre

Before filtering

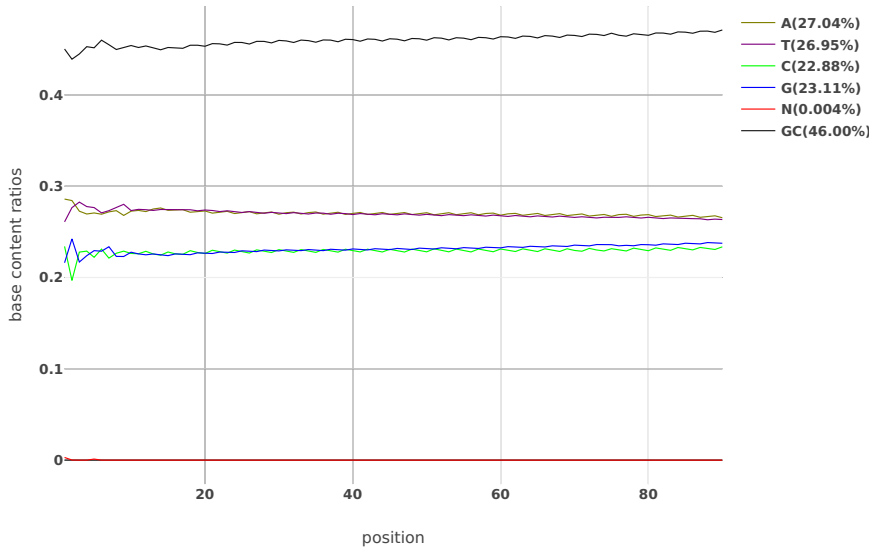
Before filtering: read1: quality

Value of each position will be shown on mouse over.



Before filtering: read1: base contents

Value of each position will be shown on mouse over.

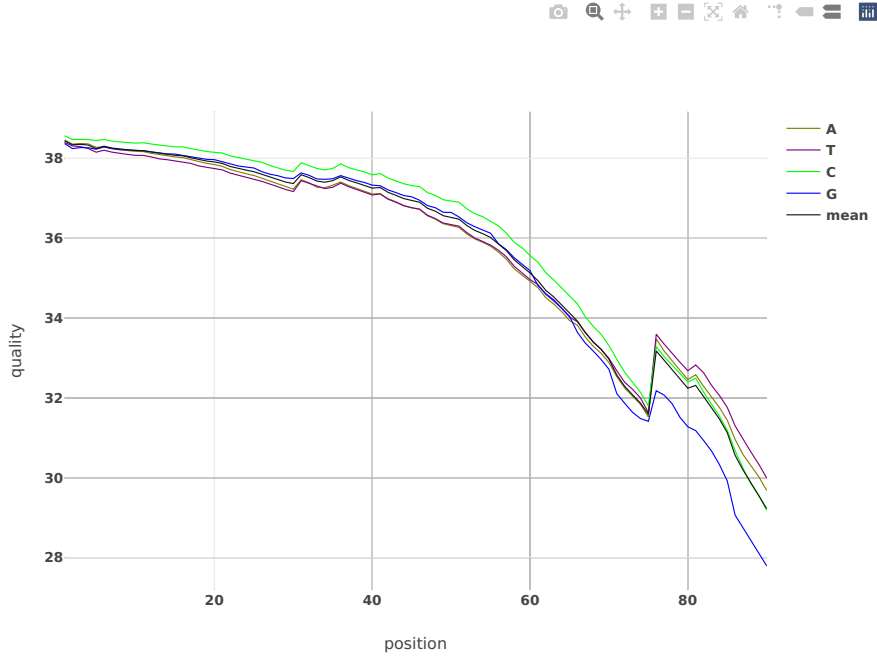


(f) Qualité, contenu en base et compte de k-mers de la lecture 1 après filtre

After filtering

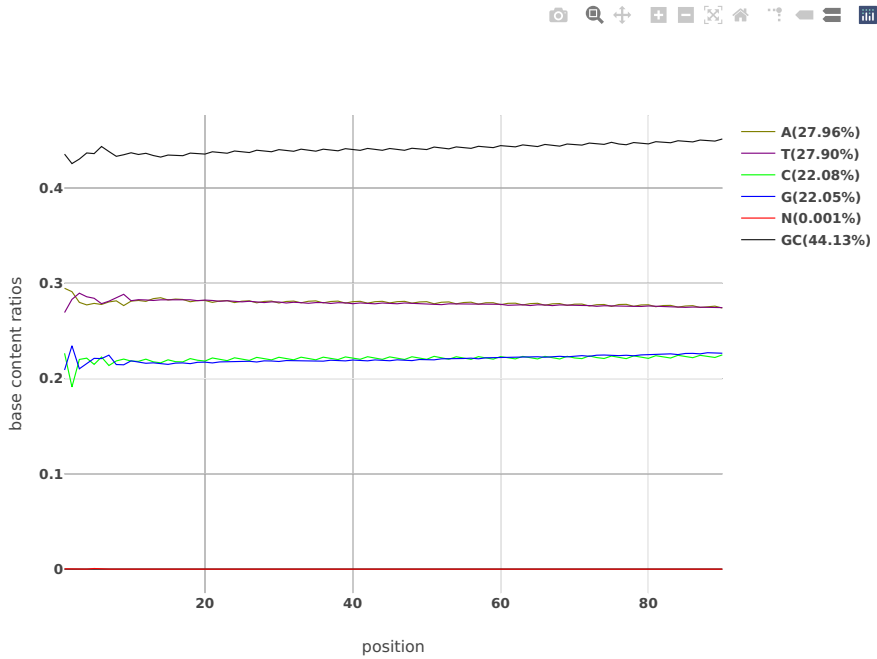
After filtering: read1: quality

Value of each position will be shown on mouse over.



After filtering: read1: base contents

Value of each position will be shown on mouse over.



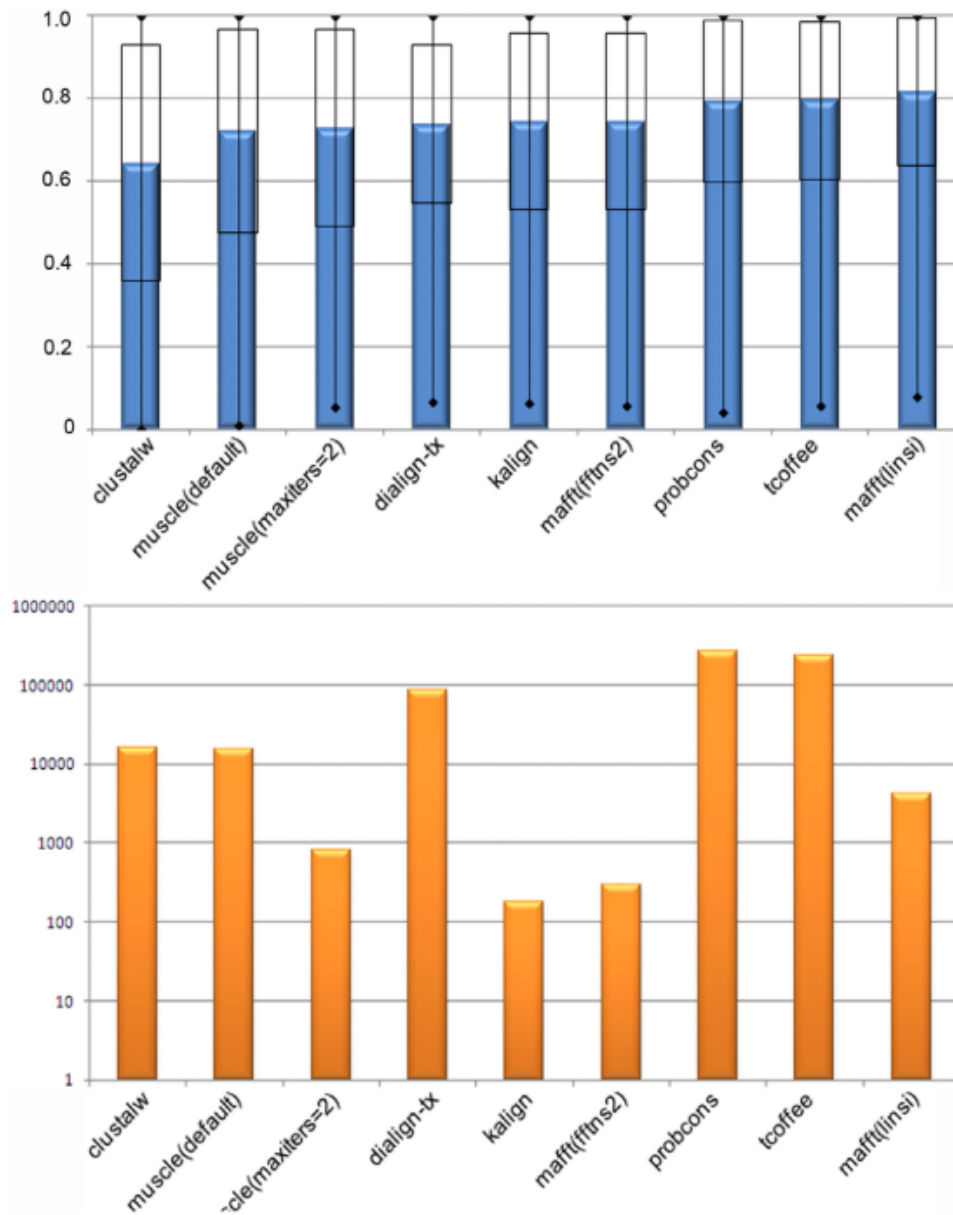


FIGURE A4 – A) Qualité de l’alignement mesuré par le score par colonne «Column Score (CS)». B) Temps d’exécution des aligneurs sur échelle logarithmique. Figure issue de Thompson et al. [388].

```

LOCUS       CP035536                185957 bp    DNA     circular BCT 10-JUL-2019
DEFINITION Klebsiella pneumoniae subsp. pneumoniae strain CCRI-22199 plasmid
            pKp199-1, complete sequence.
ACCESSION  CP035536
VERSION    CP035536.1
DBLINK     BioProject: PRJNA481972
            BioSample: SAMN09692822
KEYWORDS   .
SOURCE     Klebsiella pneumoniae subsp. pneumoniae
ORGANISM   Klebsiella pneumoniae subsp. pneumoniae
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
            Enterobacteriaceae; Klebsiella.
REFERENCE  1 (bases 1 to 185957)
AUTHORS    Deraspe,M., Longtin,J. and Roy,P.H.
TITLE      Genome sequence of a Klebsiella pneumoniae NDM-1 producer isolated
            in Quebec City
JOURNAL    Unpublished
REFERENCE  2 (bases 1 to 185957)
AUTHORS    Deraspe,M., Longtin,J. and Roy,P.H.
TITLE      Direct Submission
JOURNAL    Submitted (30-JAN-2019) Infectious Diseases Research Center,
            Universite Laval, 2705 boul. Laurier, Quebec City, Quebec G1V 4G2,
            Canada
COMMENT    Bacteria and source DNA available from Infectious Diseases Res.
            Ctr.

            ##Genome-Assembly-Data-START##
            Assembly Date      :: 16-SEP-2015
            Assembly Method    :: Celera Assembler v. SEPT-2016
            Assembly Name      :: Uvalal_Kp711_1.0
            Genome Representation :: Full
            Expected Final Version :: Yes
            Genome Coverage    :: 140.0x
            Sequencing Technology :: PacBio RSII
            ##Genome-Assembly-Data-END##
FEATURES   Location/Qualifiers
            source             1..185957
                                /organism="Klebsiella pneumoniae subsp. pneumoniae"
                                /mol_type="genomic DNA"
                                /strain="CCRI-22199"
                                /isolation_source="anal sample"
                                /host="Homo sapiens"
                                /sub_species="pneumoniae"
                                /db_xref="taxon:72407"
                                /plasmid="pKp199-1"
                                /country="Canada:Quebec City"
                                /lat_lon="46.8371 N 71.2265 W"
                                /collection_date="10-Sep-2012"
                                /collected_by="Jean Longtin"
            gene               complement(22..375)
                                /locus_tag="Kp711_6001"
            CDS                complement(22..375)
                                /locus_tag="Kp711_6001"
                                /codon_start=1
                                /transl_table=11
                                /product="copper/silver efflux system periplasmic protein
                                CusF"
                                /protein_id="QBA57780.1"
                                /translation="MRNSLKAVLFGAFSVMFSAAGLHAETHQHGDMMNAASDASVQQVIK
                                GTGVVKIDMNSKKITISHEAIPAVGWPAMTMRFTFVNADDAINALKTGNHVDFFSFIQ
                                QGNISLLKSINVTQS"
            gene               complement(404..1789)
                                /locus_tag="Kp711_6002"
            CDS                complement(404..1789)
                                /locus_tag="Kp711_6002"
                                /codon_start=1
                                /transl_table=11
                                /product="copper/silver efflux system outer membrane
                                protein CusC"
                                /protein_id="QBA57781.1"
                                /translation="MFKLLKLLSISTIFILAGCVSLAPEYQRPPAPVPQQFSLSKNSLT
                                PAVNSYQDTGWRNFFVDPQVSRLLIGEALNNRDLRMAALKVVEEARAQFNVTADADRYPQ
                                LNASSGITYNGGLKGDKPTTQEYDAGLELSYELDFFGKLNKMSADRQNYFASEEARR
                                AVHILLVSNVSQSYFSSQLLAYEQLRIARETLKNYEQSYAFVEQQLVTGSTNVLALQEA
                                RGQIESTRAEIAKREGDLAQANNALQLVLGTYRALPSEKGMKGGEIAPVKLPPNLSQ
                                ILLQRPDI MEAEYQLKAADANIGAAARAAFFPSITLTSGLSASSTELSSLFTSGSGHWN
                                FIPKIEIPFNAGRKNANLKLAEIRQQQSVVNYEQIQSAFKDVSDTLALRDSLSQL
                                ESQRYLDSLQITLQARGLYASGAVSYIEVLD AERSLFATQQTILDLTYSRVQVNEIN
                                LFTALGGGWVE"

```

[...]

FIGURE A5 – Exemple d'un fichier GenBank pour le plasmide pKp199-1 de la souche *Klebsiella pneumoniae subsp. pneumoniae* CCRI-22199 [88].

```

ID CP035536; SV 1; circular; genomic DNA; STD; PRO; 185957 BP.
XX
AC CP035536;
XX
PR Project:PRJNA481972;
XX
DT 11-FEB-2019 (Rel. 139, Created)
DT 15-JUL-2019 (Rel. 141, Last updated, Version 2)
XX
DE Klebsiella pneumoniae subsp. pneumoniae strain CCRI-22199 plasmid pKp199-1,
DE complete sequence.
XX
KW .
XX
OS Klebsiella pneumoniae subsp. pneumoniae
OC Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
OC Enterobacteriaceae; Klebsiella.
OG Plasmid pKp199-1
XX
RN [1]
RP 1-185957
RA Deraspe M., Longtin J., Roy P.H.;
RT "Genome sequence of a Klebsiella pneumoniae NDM-1 producer isolated in
RT Quebec City";
RL Unpublished.
XX
RN [2]
RP 1-185957
RA Deraspe M., Longtin J., Roy P.H.;
RT ;
RL Submitted (30-JAN-2019) to the INSDC.
RL Infectious Diseases Research Center, Universite Laval, 2705 boul. Laurier,
RL Quebec City, Quebec G1V 4G2, Canada
XX
DR MD5; e7f6a5085377fd5f42e06b54f0cd12ed.
DR BioSample; SAMN09692822.
XX
CC Bacteria and source DNA available from Infectious Diseases Res.
CC Ctr.
CC ##Genome-Assembly-Data-START##
CC Assembly Date      :: 16-SEP-2015
CC Assembly Method    :: Celera Assembler v. SEPT-2016
CC Assembly Name      :: Ulaval_Kp711_1.0
CC Genome Representation :: Full
CC Expected Final Version :: Yes
CC Genome Coverage     :: 140.0x
CC Sequencing Technology :: PacBio RSII
CC ##Genome-Assembly-Data-END##
XX
FH Key                Location/Qualifiers
FH
FH source             1..185957
FH                    /organism="Klebsiella pneumoniae subsp. pneumoniae"
FH                    /plasmid="pKp199-1"
FH                    /host="Homo sapiens"
FH                    /sub_species="pneumoniae"
FH                    /strain="CCRI-22199"
FH                    /mol_type="genomic DNA"
FH                    /country="Canada:Quebec City"
FH                    /lat_lon="46.84 N 71.23 W"
FH                    /isolation_source="anal sample"
FH                    /collected_by="Jean Longtin"
FH                    /collection_date="10-Sep-2012"
FH                    /db_xref="taxon:72407"
FH gene               complement(22..375)
FH                    /locus_tag="Kp711_6001"
FH CDS                 complement(22..375)
FH                    /codon_start=1
FH                    /transl_table=11
FH                    /locus_tag="Kp711_6001"
FH                    /product="copper/silver efflux system periplasmic protein
FH                    CusF"
FH                    /protein_id="QBA57780.1"
FH                    /translation="MRNSLKAVLFGAFSVMFSAAGLHAETHQHGMNAASDASVQQVIKG
FH                    TGVVKDIDMNSKKITISHEAIPAVGWPAMTMRFTFVNADDAINALKTGNHVDVFSFIQQG
FH                    NISLLKSINVTQS"
XX
[...]
```

FIGURE A6 – Exemple d'un fichier EMBL pour le plasmide pKp199-1 de la souche *Klebsiella pneumoniae subsp. pneumoniae* CCRI-22199 [88].

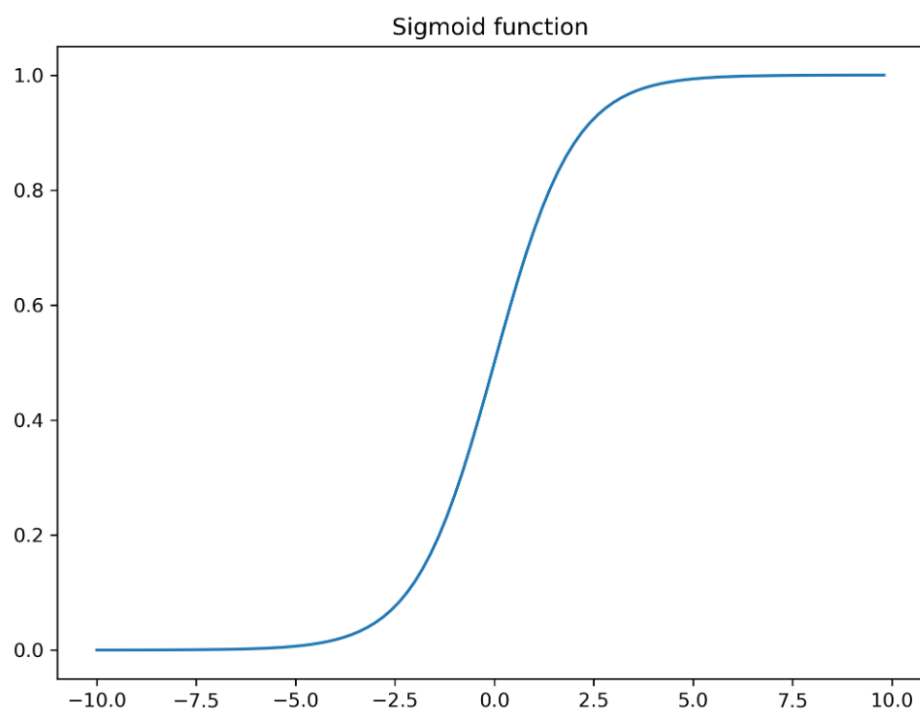


FIGURE A7 – Exemple d’une fonction logistique aussi nommé fonction sigmoïde.

```
cp035536 . gene 22 375 . - . ID=Kp711_6001
cp035536 . CDS 22 375 . - . ID=Kp711_6001; locus_tag=Kp711_6001; product=copper/silver efflux system periplasmic protein
CusF; protein_id=QBA57780.1; transL_table=11; translation=MRNSLKAVLFGAFSVMFASAGLHAETHQGDHMAASDASVQVIK GTGVVKIDIMNSKKITISHEAIPAVGVPMATMRFTFVNADDAINALKGNHVDVFSFIQ QGNISLLKSINVTQS
cp035536 . gene 404 1789 . - . ID=Kp711_6002
cp035536 . CDS 404 1789 . - . ID=Kp711_6002; locus_tag=Kp711_6002; product=copper/silver efflux system outer membrane protein
CusC; protein_id=QBA57781.1; transL_table=11; translation=MFKLKLLSISTIFILAGCVSLAPEYQRPPAPVPQQFSLSKNSLT PAVNSYQDTGWRNFVDPQVSRLIGELANNRDLRMAALKVEEARAQFNVTDADRYPP
LNASSGITVNGGLKDKPTTQEYDAGLELSVLEDFGFKLKNMSEADRNQYFASEARR AVHILLVSNSQVSFSQQLAVEQLRIARETLKNVYEQSVAFVEQQLVTG6TNVLALEQA
RGQIESITRAEIAKREGDLAQQNALLQVLTGYRALPSEKMGKEIAPVKLPPNLSQ ILLQRPDIWEAYQLKAADAMIGAARAAFPPSITLTSGLSASSTELSSLFTSGSGMMN
FITPKIEIPIFNAGRKNLKLAEIRQQSVVNYEQIKQSFAKDVSDTLALRDLSSLQ ESQQRYLDSLQITLQRAIRGLYASGAVSYIEVLDAERSLFATQQTIIIDLTVSRQVNEIN LFTALGGGWVE
cp035536 . gene 1979 2659 . + . ID=Kp711_6003
cp035536 . CDS 1979 2659 . + . ID=Kp711_6003; locus_tag=Kp711_6003; product=DNA-binding response regulator
CusR; protein_id=QBA57782.1; transL_table=11; translation=MKILLVEDEIKTGEYLSKGLTEAGFVDHADNGLTGYHLAMTAE YDLVILDIMLPDVGWDIRMRSAGKMPVLLLTALGTIEHRVKGLELGAADDYLKPV
FAFAELLARVRLRRGNMITEQLKVAADLSVDSRKRVSAGNRIVLTSKEFSLLE FFRHQGEVPLRSLIASQVWMNFSDTNAIDAVKRLRAKIDNDYGTKLITVIRGVG YMLEIPDA
cp035536 . gene 2652 4127 . + . ID=Kp711_6004
cp035536 . CDS 2652 4127 . + . ID=Kp711_6004; locus_tag=Kp711_6004; product=sensor kinase
CusS; protein_id=QBA57783.1; transL_table=11; translation=MHSKPSRRPFSLALRLTFFISLSTILAFIAFTWFMLHSVEKHFA EQDVSDLQIISTLLSRILQSPADPEKVKVSKIKESIASYRNWALLLNPRGEVLFSSA
QGAALRPVANSADPSEHRSRARDVFLWTVEDPAGPMDTGSEMKEYRIASSGQAIQF GKQQYVMYTLGSLINFHLHYLDALCKNLIAIAVVISLLVLIIIRIARVQGHPLRNVS
NAIKNITSENLDARLETRVPIELEQLVISFNHMGKIEDVFTROANFSADIAEIRT PITNLVTQTIEALSQRDTRQREDEVLKYSSLEEYNRMTKMSVDMFLAQADNNQILPDR
VNFDLRAEVMKVFEEFEAWAEEERNITLKFNGMPCLVEGDQPMFRRAINLNLALRYT PEGQAITVSIREQESFFDLVTEINPGKPIPEEHL SRLFRDRYRVDPSRQRKREGSGIGL AIKVSIVEAHRVQVESDVRSTRFRLVSPRLEKIMPETQC
cp035536 . gene 4378 4809 . + . ID=Kp711_6005
cp035536 . CDS 4378 4809 . + . ID=Kp711_6005; locus_tag=Kp711_6005; product=silver-binding protein
Si1E; protein_id=QBA57784.1; transL_table=11; translation=MKNIVLASLLGFGLLISSAWATETVNIHDRVNAQAPAHMQSAE APVGIQGTAPRMTGMDQHEQAIHAHEMTNGSADAHQMVESHQKMMGNVSTTVPS
TSYAMNHERAAVAHEFMNNGSQSHPQAMEAHRMINAG
cp035536 . gene 5105 5509 . - . ID=Kp711_6006; note=ISKpn25 transposase%2C N-terminal truncated
cp035536 . gene 5663 6040 . - . ID=Kp711_6007; note=IS1 OrfB%2C N-terminal truncated
cp035536 . gene 6211 6606 . + . ID=Kp711_6008
cp035536 . CDS 6211 6606 . + . ID=Kp711_6008; locus_tag=Kp711_6008; product=ISEc11 transposase%2C
OrfA; protein_id=QBA57785.1; transL_table=11; translation=MSNTNANFEMTGLLQGEARKRKTPEKIAIIQQTMEPGMVSH VARLHGIQPSLLFKWKKYQEGSLTAVAAGEEVPASELTAALKQVRELQRLGKKTM
EVEILKEAVEYGSRKWAHAPLPLPKDGE
cp035536 . gene 6564 7445 . + . ID=Kp711_6009
cp035536 . CDS 6564 7445 . + . ID=Kp711_6009; locus_tag=Kp711_6009; product=ISEc11-like transposase%2C
OrfB; protein_id=QBA57786.1; transL_table=11; translation=MDSARALVAKGRIALVSRMTGSRAQLSLRINRSADWDKRCN RRNDEADEILSAILDIIISDMPYSYGRVWGLLRKQRTREGQPPVNAKRLYRIMSEHN
LLLLHDKPERPKREHKGIAVAESDMRWCSDGFEGCDNGEKL RVTFALDCCDREAID WAASTGGYDSSVTQDVMRLSVEKRFGRDLPTAVQWLTDNGSAYTAETWRFARELNL
EPCITAVSSPQNGMAERFVKTKEDYIAFMKPDVRTALRNLVAFTHYENHPSA LGYHSPREYRRQRTSLT
cp035536 . gene 7830 8333 . + . ID=Kp711_6010
cp035536 . CDS 7830 8333 . + . ID=Kp711_6010; locus_tag=Kp711_6010; product=IS1 transposase
OrfB; protein_id=QBA57787.1; transL_table=11; translation=MSRQCTHYGRWPGHGFTSLKKLRPQSVTSRIQP6SDVIICAEID EQWGYVGAKSRQRWLFAYDRIRRTVVAVHVGERTLATLERLLSLSAFEVVMMDTG
WPLYESRLKGLKHVISKRYQRIERHNLNRQHLARLGRKSLFSKSVELYDKVIGHY LNIKHYQ
cp035536 . gene 8485 10482 . + . ID=Kp711_6011
cp035536 . CDS 8485 10482 . + . ID=Kp711_6011; locus_tag=Kp711_6011; product=choline
transporter; protein_id=QBA57788.1; transL_table=11; translation=MSDNDTIPKKSQINXAVFFTSALLIFLLVAFVAAPVDDADKN FKLLQQQIFTNNAWFIYILAVALLISLVTLFLGRSRYDGLKGPDHAQPDFSYHSWFAML
FSAGMGIQLMFFGVAEPVMHYLSPPPVGTPEVAAAKEAMRLTFFHMGHAWAIYAIVA LILAFFSYRHGLPLTLRSALYPIGDRIYGPVGHAVDIFAVIGTVFVATSLGTYGLVQ
VNAGLNHLFGVPIINETVQVLLIVITGLATISVVSGLKGRIRLSLNLGLALLLALL VLCLGPTVLNLSKSFENTGGVLSLVSKTFNLYAYEPKSSKMLGWTLLYWGWL5W5
PFYGVGFIARVSRGRTRREFVTGVLFPAGFTLMmMTVFGNSATYIMNQGATDLANTV QQDVALALFNFLEHFPFSSVLSFIAMAMVIVFVTSADSGAMVVDTLATGGVANTPW
QRFIWASLMGIVATLALLAGGSLAQVTVITASALPFSVILLISIVGLKALRRDLTKR ESLSMATIAPTAARNPIPWQRRLRNIAYLKRSLLVKRFBMDVIVQPMTLVQEELNKQ
TISHISDAVEDRIRLEVDQLNELNFYEVRLRGYSSPTFALAAMDNEQTEQHRYR AEVYKKEGGQNYDVMGNQEQLINDLDQYEKHLHFLHLVR
cp035536 . gene 10545 11822 . + . ID=Kp711_6012
cp035536 . CDS 10545 11822 . + . ID=Kp711_6012; locus_tag=Kp711_6012; product=hypothetical
protein; protein_id=QBA57789.1; transL_table=11; translation=MISRWKWLKQTKLWFRATLFAIVAITALLSILFKSMIPES VSKVGAEAVDNIINLASMLAVTTFSLSMVTAYGSAATNVTPRATRLVEDVTQQ
NVLATIFGSLFSLVIGIALSMDGAYGERGRVILFIVTLVIALIITLLRWIQLTSL GRVGETTAKVEQAETTFARARNPCLGGYVWLENNEOPKGTAVVYPKIGYVEYINM
VAKLSKLLTNDPRHYLVLAQPGSFHPMPVLYL SQGQESSISADLLETIIVDVRSA QDPRFCLVMAEACRALSVAVDNPGTADIVIGRVIRLSTYAQKNSDEIYKVP5VH
VPLQNDLLEDFFSPVARDGAGMREIQIRVLKGLSMLSKWGPGFSEAAHNAFALET EHAIRADHIDSDRCLIKLYYNLFSGEDSNKKP
[...]
```

FIGURE A8 – Exemple d'un fichier GFF pour le plasmide pKp199-1 de la souche *Klebsiella pneumoniae subsp. pneumoniae* CCRI-22199 [88].

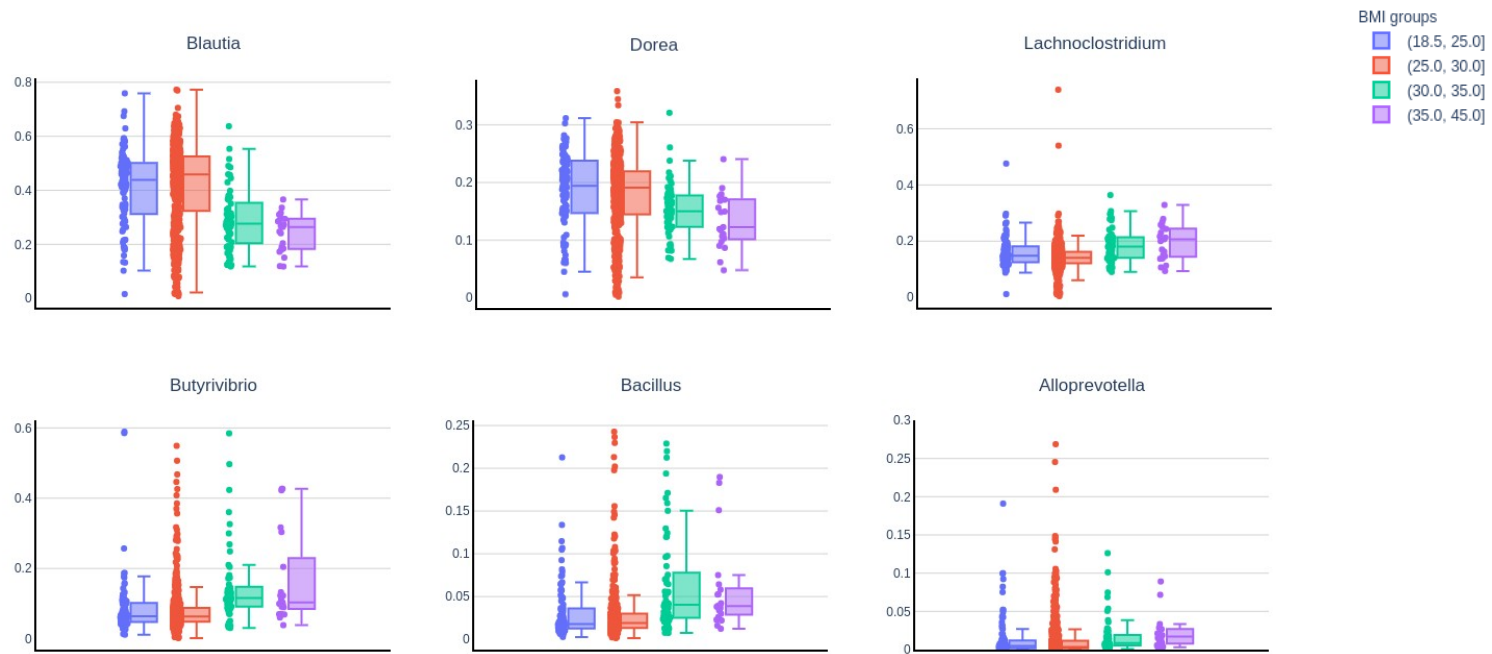


FIGURE A9 – The top 6 most abundant genera with significant changes in obese and non-obese gut microbiota.

Tableaux

TABLEAU A1 – Bases de données utilisées en génomique bactérienne

Base de données	Description	Spécialité	Référence	URL
CARD	The Comprehensive Antibiotic Resistance Database	Résistance	[258; 183; 5]	https://card.mcmaster.ca/
ResFinder	ResFinder identifies acquired antimicrobial resistance genes and/or chromosomal mutations in total or partial sequenced isolates of bacteria	Résistance	[435]	https://cge.cbs.dtu.dk/services/ResFinder/
ARG-ANNOT	Antibiotic Resistance Gene ANNOTation	Résistance	[153]	https://www.mediterranee-infection.com/acces-ressources/base-de-donnees/arg-annot-2/
ARDB	Antibiotic Resistance Genes Database	Résistance	[237]	https://ardb.cbc.umd.edu/
VFDB	Virulence Factor Database	Virulence	[56; 429; 55; 57; 238]	http://www.mgc.ac.cn/VFs/
PHI-base	The Pathogen-Host Interactions Database	Virulence	[415; 416; 397; 395; 396]	http://www.phi-base.org/
Victors	Victors is a manually curated, web-based integrative knowledge base and analysis resource for VFs of pathogens	Virulence	[343]	http://www.phidias.us/victors/

ISfinder	The Reference Centre for Bacterial Insertion Sequences (IS)	Éléments mobiles	[360]	https://isfinder.biotoul.fr/
ICEberg	Integrative and Conjugative Elements found in Bacteria	Éléments mobiles	[29; 240]	https://db-mml.sjtu.edu.cn/ICEberg/
Integrall	Database and search engine for integrons, integrases and gene cassettes	Éléments mobiles	[268]	http://integrall.bio.ua.pt/
PhagesFinder	Automated identification and classification of prophage regions in complete bacterial genome sequences	Bactériophages	[122]	http://phage-finder.sourceforge.net/
PHAST / PHASTER	A Fast Phage Search Tool	Bactériophages	[441; 16]	http://phast.wishartlab.com/index.html ; https://phaster.ca/
KEGG	Kyoto Encyclopedia of Genes and Genomes	Sentiers métabolique	[285; 190; 192; 191]	https://www.genome.jp/kegg/
BioCyc / EcoCyc / MetaCyc	Metabolic Pathways of microbial, <i>Escherichia coli</i> and all domain of live	Sentiers métabolique	[196; 195; 197]	https://biocyc.org/ ; https://metacyc.org/ ; https://ecocyc.org/
PubMLST	Public databases for molecular typing and microbial genome diversity	Générique	[187]	https://pubmlst.org/
COG	The database of Clusters of Orthologous Groups of proteins	Générique	[381; 130]	https://www.ncbi.nlm.nih.gov/COG/
Gene Ontology	The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes	Générique	[18; 386]	http://geneontology.org/

eggNOG	A database of orthology relationships, functional annotation, and gene evolutionary histories	Générique	[182; 172]	http://eggnog5.embl.de/
NCBI RefSeq	A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein	Générique	[383; 382; 287]	https://www.ncbi.nlm.nih.gov/refseq/
EBI Uniprot	The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information	Générique	[387]	https://www.uniprot.org/

Bibliographie

- [1] 1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., and McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422) :56–65.
- [2] Ahmadi, A., Behm, A., Honnalli, N., Li, C., Weng, L., and Xie, X. (2012). Hobbes : optimized gram-based methods for efficient read alignment. *Nucleic acids research*, 40(6) :e41.
- [3] Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society : Series B (Methodological)*, 44(2) :139–160.
- [4] Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna : A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- [5] Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A.-L. V., Cheng, A. A., Liu, S., Min, S. Y., Miroshnichenko, A., Tran, H.-K., Werfalli, R. E., Nasir, J. A., Oloni, M., Speicher, D. J., Florescu, A., Singh, B., Faltyn, M., Hernandez-Koutoucheva, A., Sharma, A. N., Bordeleau, E., Pawlowski, A. C., Zubyk, H. L., Dooley, D., Griffiths, E., Maguire, F., Winsor, G. L., Beiko, R. G., Brinkman, F. S. L., Hsiao, W. W. L., Domselaar, G. V., and McArthur, A. G. (2019). CARD 2020 : antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Research*.
- [6] Algoblan, A., Alalfi, M., and Khan, M. (2014). Mechanism linking diabetes mellitus and obesity. *Diabetes, Metabolic Syndrome and Obesity : Targets and Therapy*, page 587.
- [7] Allison, G. E. and Verma, N. K. (2000). Serotype-converting bacteriophages and O-antigen modification in *Shigella flexneri*. *Trends in microbiology*, 8(1) :17–23.
- [8] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3) :403–10.
- [9] Alves, A., Bassot, A., Bulteau, A.-L., Pirola, L., and Morio, B. (2019). Glycine Metabolism and Its Alterations in Obesity and Metabolic Diseases. *Nutrients*, 11(6) :1356.

- [10] Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., Kightley, E. P., Thompson, L. R., Hyde, E. R., Gonzalez, A., and Knight, R. (2017). Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems*, 2(2).
- [11] Amrane, S., Hocquart, M., Afouda, P., Kuete, E., Pham, T.-P.-T., Dione, N., Ngom, I. I., Valles, C., Bachar, D., Raoult, D., and Lagier, J. C. (2019). Metagenomic and culturomic analysis of gut microbiota dysbiosis during *Clostridium difficile* infection. *Scientific Reports*, 9(1) :12807.
- [12] Andam, C. P. and Hanage, W. P. (2015). Mechanisms of genome evolution of *Streptococcus*. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 33 :334–42.
- [13] Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral ecology*, 26(1) :32–46.
- [14] Andres, E. (2004). Vitamin B12 (cobalamin) deficiency in elderly patients. *Canadian Medical Association Journal*, 171(3) :251–259.
- [15] Arboleya, S., Watkins, C., Stanton, C., and Ross, R. P. (2016). Gut Bifidobacteria Populations in Human Health and Aging. *Frontiers in Microbiology*, 7.
- [16] Arndt, D., Grant, J. R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart, D. S. (2016). PHASTER : a better, faster version of the PHAST phage search tool. *Nucleic Acids Research*, 44(W1) :W16–W21.
- [17] Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borrueal, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H. B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E. G., Wang, J., Guarner, F., Pedersen, O., de Vos, W. M., Brunak, S., Doré, J., Weissenbach, J., Ehrlich, S. D., and Bork, P. (2011). Enterotypes of the human gut microbiome. *Nature*, 473(7346) :174–180.
- [18] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology : tool for the unification of biology. *Nature Genetics*, 25(1) :25–29.
- [19] Ayling, M., Clark, M. D., and Leggett, R. M. (2020). New approaches for metagenome assembly with short reads. *Briefings in Bioinformatics*, 21(2) :584–594.

- [20] Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., and Zagnitko, O. (2008). The RAST Server : Rapid Annotations using Subsystems Technology. *BMC Genomics*, 9(1) :75.
- [21] Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Research*, 28(1) :304–305.
- [22] Bajzer, M. and Seeley, R. J. (2006). Obesity and gut flora. *Nature*, 444(7122) :1009–1010.
- [23] Balvočiūtė, M. and Huson, D. H. (2017). SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare? *BMC genomics*, 18(Suppl 2) :114.
- [24] Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012). SPAdes : A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5) :455–477.
- [25] Baquero, F. and Nombela, C. (2012). The microbiome as a human organ. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 18 Suppl 4 :2–4.
- [26] Barker, H. A. (1981). Amino Acid Degradation by Anaerobic Bacteria. *Annual Review of Biochemistry*, 50(1) :23–40.
- [27] Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M.-C. C., Charles, T., Chen, X., Cocolin, L., Eversole, K., Corral, G. H., Kazou, M., Kinkel, L., Lange, L., Lima, N., Loy, A., Macklin, J. A., Maguin, E., Mauchline, T., McClure, R., Mitter, B., Ryan, M., Sarand, I., Smidt, H., Schelkle, B., Roume, H., Kiran, G. S., Selvin, J., de Souza, R. S. C., van Overbeek, L., Singh, B. K., Wagner, M., Walsh, A., Sessitsch, A., and Schloter, M. (2020). Microbiome definition re-visited : old concepts and new challenges. *Microbiome*, 8(1) :103.
- [28] Besemer, J. (2001). GeneMarkS : a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research*, 29(12) :2607–2618.
- [29] Bi, D., Xu, Z., Harrison, E. M., Tai, C., Wei, Y., He, X., Jia, S., Deng, Z., Rajakumar, K., and Ou, H.-Y. (2012). ICEberg : a web-based resource for integrative and conjugative elements found in Bacteria. *Nucleic Acids Research*, 40(D1) :D621–D626.
- [30] Biek, R., Pybus, O. G., Lloyd-Smith, J. O., and Didelot, X. (2015). Measurably evolving pathogens in the genomic era. *Trends in ecology & evolution*, 30(6) :306–13.

- [31] Boc, A., Diallo, A. B., and Makarenkov, V. (2012). T-REX : a web server for inferring, validating and visualizing phylogenetic trees and networks. *Nucleic acids research*, 40(Web Server issue) :W573–9.
- [32] Boffa, L. C., Vidali, G., Mann, R. S., and Allfrey, V. G. (1978). Suppression of histone deacetylation in vivo and in vitro by sodium butyrate. *Journal of Biological Chemistry*, 253(10) :3364–3366.
- [33] Boisvert, S., Laviolette, F., and Corbeil, J. (2010). Ray : simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of computational biology : a journal of computational molecular cell biology*, 17(11) :1519–1533.
- [34] Boisvert, S., Raymond, F., Godzaridis, É., Laviolette, F., and Corbeil, J. (2012). Ray Meta : Scalable de novo metagenome assembly and profiling. *Genome Biology*, 13(12) :R122.
- [35] Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic : a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, 30(15) :2114–20.
- [36] Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., Cope, E. K., Da Silva, R., Diener, C., Dorrestein, P. C., Douglas, G. M., Durall, D. M., Duvall, C., Edwardson, C. F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J. M., Gibbons, S. M., Gibson, D. L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G. A., Janssen, S., Jarmusch, A. K., Jiang, L., Kaehler, B. D., Kang, K. B., Keefe, C. R., Keim, P., Kelley, S. T., Knights, D., Koester, I., Kosciulek, T., Kreps, J., Langille, M. G. I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B. D., McDonald, D., McIver, L. J., Melnik, A. V., Metcalf, J. L., Morgan, S. C., Morton, J. T., Naimey, A. T., Navas-Molina, J. A., Nothias, L. F., Orchanian, S. B., Pearson, T., Peoples, S. L., Petras, D., Preuss, M. L., Pruesse, E., Rasmussen, L. B., Rivers, A., Robeson, M. S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S. J., Spear, J. R., Swafford, A. D., Thompson, L. R., Torres, P. J., Trinh, P., Tripathi, A., Turnbaugh, P. J., Ul-Hasan, S., van der Hooft, J. J. J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K. C., Williamson, C. H. D., Willis, A. D., Xu, Z. Z., Zaneveld, J. R., Zhang, Y., Zhu, Q., Knight, R., and Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, 37(8) :852–857.
- [37] Bose, T., Haque, M. M., Reddy, C., and Mande, S. S. (2015). COGNIZER : A Framework for Functional Annotation of Metagenomic Datasets. *PLOS ONE*, 10(11) :e0142102.

- [38] Botzman, M. and Margalit, H. (2011). Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome biology*, 12(10) :R109.
- [39] Braslavsky, I., Hebert, B., Kartalov, E., and Quake, S. R. (2003). Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences*, 100(7) :3960–3964.
- [40] Bray, J. R. and Curtis, J. T. (1957). An ordination of upland forest communities of southern Wisconsin. *Ecological Monographs* (27).
- [41] Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.
- [42] Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., Olson, R., Overbeek, R., Parrello, B., Pusch, G. D., Shukla, M., Thomason, J. A., Stevens, R., Vonstein, V., Wattam, A. R., and Xia, F. (2015). RASTtk : a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific reports*, 5 :8365.
- [43] Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature methods*, 12(1) :59–60.
- [44] Burkov, A. (2019). *The Hundred-Page Machine Learning Book*. Andriy Burkov.
- [45] Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2 : High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13(7) :581–583.
- [46] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST+ : architecture and applications. *BMC bioinformatics*, 10 :421.
- [47] Candido, E. P. M., Reeves, R., and Davie, J. R. (1978). Sodium butyrate inhibits histone deacetylation in cultured cells. *Cell*, 14(1) :105–113.
- [48] Canzar, S. and Salzberg, S. L. (2017). Short Read Mapping : An Algorithmic Tour. *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers*, 105(3) :436–458.
- [49] Castro-Nallar, E., Bendall, M. L., Pérez-Losada, M., Sabuncyan, S., Severance, E. G., Dickerson, F. B., Schroeder, J. R., Yolken, R. H., and Crandall, K. A. (2015). Composition, taxonomy and functional diversity of the oropharynx microbiome in individuals with schizophrenia and controls. *PeerJ*, 3 :e1140.
- [50] CDC (2020). Defining Adult Overweight and Obesity (<https://www.cdc.gov/obesity/adult/defining.html>).

- [51] Chambers, E. S., Preston, T., Frost, G., and Morrison, D. J. (2018). Role of Gut Microbiota-Generated Short-Chain Fatty Acids in Metabolic and Cardiovascular Health. *Current Nutrition Reports*, 7(4) :198–206.
- [52] Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., Chandra, T., Fikes, A., and Gruber, R. E. (2008). Bigtable. *ACM Transactions on Computer Systems*, 26(2) :1–26.
- [53] Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of statistics*, pages 265–270.
- [54] Check Hayden, E. (2012). Nanopore genome sequencer makes its debut. *Nature*.
- [55] Chen, L., Xiong, Z., Sun, L., Yang, J., and Jin, Q. (2012). VFDB 2012 update : toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic acids research*, 40(Database issue) :D641–5.
- [56] Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., and Jin, Q. (2005). VFDB : a reference database for bacterial virulence factors. *Nucleic acids research*, 33(Database issue) :D325–8.
- [57] Chen, L., Zheng, D., Liu, B., Yang, J., and Jin, Q. (2016). VFDB 2016 : Hierarchical and refined dataset for big data analysis - 10 years on. *Nucleic Acids Research*, 44(D1) :D694–D697.
- [58] Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp : an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17) :i884–i890.
- [59] Chen, T. and Guestrin, C. (2016). XGBoost : A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.
- [60] Chowdhury, B. and Garai, G. (2017). A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics*, 109(5-6) :419–431.
- [61] Christensen, H. and Moodley, A. (2018). *Introduction to Bioinformatics in Microbiology*. Learning Materials in Biosciences. Springer International Publishing, Cham.
- [62] Chun, J., Oren, A., Ventosa, A., Christensen, H., Arahal, D. R., da Costa, M. S., Rooney, A. P., Yi, H., Xu, X.-W., De Meyer, S., and Others (2018). Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *International journal of systematic and evolutionary microbiology*, 68(1) :461–466.
- [63] Cimermancic, P., Medema, M. H., Claesen, J., Kurita, K., Wieland Brown, L. C., Mavrommatis, K., Pati, A., Godfrey, P. A., Koehrsen, M., Clardy, J., Birren, B. W., Takano, E.,

- Sali, A., Linington, R. G., and Fischbach, M. A. (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, 158(2) :412–21.
- [64] Clarke, E. L., Taylor, L. J., Zhao, C., Connell, A., Lee, J.-J., Fett, B., Bushman, F. D., and Bittinger, K. (2019). Sunbeam : an extensible pipeline for analyzing metagenomic sequencing experiments. *Microbiome*, 7(1) :46.
- [65] Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian journal of ecology*, 18(1) :117–143.
- [66] Cobb, M. (2017). 60 years ago, Francis Crick changed the logic of biology. *PLOS Biology*, 15(9) :e2003243.
- [67] Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. (2009). Biopython : freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics (Oxford, England)*, 25(11) :1422–3.
- [68] Cohen, S. N., Chang, A. C., Boyer, H. W., and Helling, R. B. (1973). Construction of biologically functional bacterial plasmids in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, 70(11) :3240–4.
- [69] Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., Brown, C. T., Porras-Alfaro, A., Kuske, C. R., and Tiedje, J. M. (2014). Ribosomal Database Project : data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, 42(D1) :D633–D642.
- [70] Collaborators, T. G. . O. (2017). Health Effects of Overweight and Obesity in 195 Countries over 25 Years. *New England Journal of Medicine*, 377(1) :13–27.
- [71] Colombo, M.-L., Hanique, S., Baurin, S. L., Bauvois, C., De Vriendt, K., Van Beeumen, J. J., Frère, J.-M., and Joris, B. (2004). The ybxI gene of *Bacillus subtilis* 168 encodes a class D beta-lactamase of low activity. *Antimicrobial agents and chemotherapy*, 48(2) :484–90.
- [72] Compeau, P. E. C., Pevzner, P. A., and Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, 29(11) :987–991.
- [73] Consortium, I. H. G. S. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822) :860–921.
- [74] CRICK, F. H. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology*, 12 :138–63.
- [75] Croucher, N. J., Finkelstein, J. A., Pelton, S. I., Mitchell, P. K., Lee, G. M., Parkhill, J., Bentley, S. D., Hanage, W. P., and Lipsitch, M. (2013). Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nature genetics*, 45(6) :656–63.

- [76] Croucher, N. J., Finkelstein, J. A., Pelton, S. I., Parkhill, J., Bentley, S. D., Lipsitch, M., and Hanage, W. P. (2015). Population genomic datasets describing the post-vaccine evolutionary epidemiology of *Streptococcus pneumoniae*. *Scientific data*, 2 :150058.
- [77] Dashwood, R. H., Myzak, M. C., and Ho, E. (2006). Dietary HDAC inhibitors : time to rethink weak ligands in cancer chemoprevention ? *Carcinogenesis*, 27(2) :344–349.
- [78] Davis, J. J., Boisvert, S., Brettin, T., Kenyon, R. W., Mao, C., Olson, R., Overbeek, R., Santerre, J., Shukla, M., Wattam, A. R., Will, R., Xia, F., and Stevens, R. (2016). Antimicrobial Resistance Prediction in PATRIC and RAST. *Scientific Reports*, 6(1) :27930.
- [79] Dayhoff, M., Schwartz, R., and Orcutt, B. (1978). 22 a model of evolutionary change in proteins. In *Atlas of protein sequence and structure*, volume 5, pages 345–352. National Biomedical Research Foundation Silver Spring MD.
- [80] Debnath, B., Sengupta, S., and Li, J. (2011). SkimpyStash. In *Proceedings of the 2011 international conference on Management of data - SIGMOD '11*, page 25, New York, New York, USA. ACM Press.
- [81] Degroeve, S., De Baets, B., de Peer, Y., and Rouzé, P. (2002). Feature subset selection for splice site prediction. *Bioinformatics*, 18(suppl_2) :S75—S83.
- [82] Denisov, G., Walenz, B., Halpern, A. L., Miller, J., Axelrod, N., Levy, S., and Sutton, G. (2008). Consensus generation and variant detection by Celera Assembler. *Bioinformatics*, 24(8) :1035–1040.
- [83] Deorowicz, S., Kokot, M., Grabowski, S., and Debudaj-Grabysz, A. (2015). KMC 2 : fast and resource-frugal k-mer counting. *Bioinformatics*, 31(10) :1569–1576.
- [84] Déraspe, M. (2015). Développement d’une base de données sur la résistance aux antibiotiques et son utilisation en génomique. *Mémoire Université Laval*.
- [85] Déraspe, M., Binkley, G., Butano, D., Chadwick, M., Cherry, J. M., Clark-Casey, J., Contrino, S., Corbeil, J., Heimbach, J., Karra, K., and Others (2016). Making Linked Data SPARQL with the InterMine Biological Data Warehouse. In *SWAT4LS*.
- [86] Déraspe, M., Boisvert, S., Laviolette, F., Roy, P. H., and Corbeil, J. (2020). Fast protein database as a service with kAAmer. *bioRxiv*.
- [87] Déraspe, M., Karra, K., Binkley, G., Sullivan, J., Micklem, G., Corbeil, J., Cherry, J. M., and Dumontier, M. (2016). Semantic Research Platform for Model Organism Data. In *Biomedical Data Integration and Discovery*.
- [88] Déraspe, M., Longtin, J., and Roy, P. H. (2020). Genome Sequence of a *Klebsiella pneumoniae* NDM-1 Producer Isolated in Quebec City. *Microbiology Resource Announcements*, 9(3).

- [89] Déraspe, M., Raymond, F., Boisvert, S., Culley, A., Roy, P. H., Laviolette, F., and Corbeil, J. (2017). Phenetic Comparison of Prokaryotic Genomes Using k-mers. *Molecular Biology and Evolution*, 34(10) :2716–2729.
- [90] Déraspe, M., Raymond, F., Boisvert, S., Culley, A., Roy, P. H., Laviolette, F., and Corbeil, J. (2017). Phenetic Comparison of Prokaryotic Genomes Using k-mers. *Molecular biology and evolution*, 34(10) :2716–2729.
- [91] DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology*, 72(7) :5069–5072.
- [92] Dgraph Labs (2017). Badger.
- [93] Di Cara, F., Andreoletti, P., Trompier, D., Vejux, A., Bülow, M. H., Sellin, J., Lizard, G., Cherkaoui-Malki, M., and Savary, S. (2019). Peroxisomes in Immune Response and Inflammation. *International Journal of Molecular Sciences*, 20(16) :3877.
- [94] Di Cara, F., Bülow, M. H., Simmonds, A. J., and Rachubinski, R. A. (2018). Dysfunctional peroxisomes compromise gut structure and host defense by increased cell death and Tor-dependent autophagy. *Molecular biology of the cell*, 29(22) :2766–2783.
- [95] Dobrindt, U., Hochhut, B., Hentschel, U., and Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nature reviews. Microbiology*, 2(5) :414–24.
- [96] Donati, C., Hiller, N. L., Tettelin, H., Muzzi, A., Croucher, N. J., Angiuoli, S. V., Oggioni, M., Dunning Hotopp, J. C., Hu, F. Z., Riley, D. R., Covacci, A., Mitchell, T. J., Bentley, S. D., Kilian, M., Ehrlich, G. D., Rappuoli, R., Moxon, E. R., and Masignani, V. (2010). Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome biology*, 11(10) :R107.
- [97] Donohoe, D. R., Garge, N., Zhang, X., Sun, W., O’Connell, T. M., Bunger, M. K., and Bultman, S. J. (2011). The Microbiome and Butyrate Regulate Energy Metabolism and Autophagy in the Mammalian Colon. *Cell Metabolism*, 13(5) :517–526.
- [98] Drouin, A., Giguère, S., Déraspe, M., Marchand, M., Tyers, M., Loo, V. G., Bourgault, A.-M., Laviolette, F., and Corbeil, J. (2016). Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC genomics*, 17(1) :754.
- [99] Duan, J., Meng, X., Liu, S., Zhou, P., Zeng, C., Fu, C., Dou, Q., Wu, A., and Li, C. (2020). Gut Microbiota Composition Associated With *Clostridium difficile*-Positive Diarrhea and *C. difficile* Type in ICU Patients. *Frontiers in Cellular and Infection Microbiology*, 10.

- [100] Duncan, S. H., Lobley, G. E., Holtrop, G., Ince, J., Johnstone, A. M., Louis, P., and Flint, H. J. (2008). Human colonic microbiota associated with diet, obesity and weight loss. *International Journal of Obesity*, 32(11) :1720–1724.
- [101] Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis*. Cambridge University Press.
- [102] Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, 14(9) :755–763.
- [103] Edgar, R. C. (2004). MUSCLE : a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5(1) :113.
- [104] Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19) :2460–2461.
- [105] Editor (2011). Outbreak genomics. *Nature Biotechnology*, 29(9) :769.
- [106] El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., and Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1) :D427–D432.
- [107] ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414) :57–74.
- [108] Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G., and Sogin, M. L. (2013). Oligotyping : differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution*, 4(12) :1111–1119.
- [109] Facebook (2013). RocksDB.
- [110] Farris, J. S. (1970). Methods for Computing Wagner Trees. *Systematic Biology*, 19(1) :83–92.
- [111] Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic acids research*, 40(Database issue) :D136–43.
- [112] Feldgarden, M., Brover, V., Haft, D. H., Prasad, A. B., Slotta, D. J., Tolstoy, I., Tyson, G. H., Zhao, S., Hsu, C.-H., McDermott, P. F., Tadesse, D. A., Morales, C., Simmons, M., Tillman, G., Wasilenko, J., Folster, J. P., and Klimke, W. (2019). Validating the AMR-Finder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrobial agents and chemotherapy*, 63(11).
- [113] Felsenstein, J. (1981). Evolutionary trees from DNA sequences : A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6) :368–376.

- [114] Feng, D.-F. and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution*, 25(4) :351–360.
- [115] Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., Xia, H., Xu, X., Jie, Z., Su, L., Li, X., Li, X., Li, J., Xiao, L., Huber-Schönauer, U., Niederseer, D., Xu, X., Al-Aama, J. Y., Yang, H., Wang, J., Kristiansen, K., Arumugam, M., Tilg, H., Datz, C., and Wang, J. (2015). Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nature communications*, 6 :6528.
- [116] Feng, X., Jin, H., Zheng, R., Zhu, L., and Dai, W. (2015). Accelerating Smith-Waterman Alignment of Species-Based Protein Sequences on GPU. *International Journal of Parallel Programming*, 43(3) :359–380.
- [117] Fenselau, C., Havey, C., Teerakulkittipong, N., Swatkoski, S., Laine, O., and Edwards, N. (2008). Identification of beta-lactamase in antibiotic-resistant *Bacillus cereus* spores. *Applied and environmental microbiology*, 74(3) :904–6.
- [118] Ferragina, P. and Manzini, G. (2000). Opportunistic data structures with applications. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 390–398. IEEE.
- [119] Fitch, W. M. (1971). Toward Defining the Course of Evolution : Minimum Change for a Specific Tree Topology. *Systematic Zoology*, 20(4) :406.
- [120] Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., and Merrick, J. M. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science (New York, N.Y.)*, 269(5223) :496–512.
- [121] Foerstner, K. U., von Mering, C., Hooper, S. D., and Bork, P. (2005). Environments shape the nucleotide composition of genomes. *EMBO reports*, 6(12) :1208–13.
- [122] Fouts, D. E. (2006). Phage_Finder : Automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Research*, 34(20) :5839–5851.
- [123] Fowlkes, E. B. and Mallows, C. L. (1983). A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383) :553.
- [124] Franzosa, E. A., Huang, K., Meadow, J. F., Gevers, D., Lemon, K. P., Bohannon, B. J. M., and Huttenhower, C. (2015). Identifying personal microbiomes using metagenomic codes. *Proceedings of the National Academy of Sciences*, 112(22) :E2930–E2938.
- [125] Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., Knight, R., Caporaso, J. G., Segata, N., and Huttenhower, C. (2018).

- Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods*, 15(11) :962–968.
- [126] Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., Fritchman, J. L., Weidman, J. F., Small, K. V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T. R., Saudek, D. M., Phillips, C. A., Merrick, J. M., Tomb, J.-F., Dougherty, B. A., Bott, K. F., Hu, P.-C., and Lucier, T. S. (1995). The Minimal Gene Complement of *Mycoplasma genitalium*. *Science*, 270(5235) :397–404.
- [127] Freund, Y. and Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1) :119–139.
- [128] Friedman, J. and Alm, E. J. (2012). Inferring Correlation Networks from Genomic Survey Data. *PLoS Computational Biology*, 8(9) :e1002687.
- [129] Galili, T. (2015). dendextend : an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics (Oxford, England)*, 31(22) :3718–20.
- [130] Galperin, M. Y., Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research*, 43(D1) :D261–D269.
- [131] Garcea, R. L. and Alberts, B. M. (1980). Comparative studies of histone acetylation in nucleosomes, nuclei, and intact cells. Evidence for special factors which modify acetylase action. *Journal of Biological Chemistry*, 255(23) :11454–11463.
- [132] Gardner, S. N., Slezak, T., and Hall, B. G. (2015). kSNP3.0 : SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics*, 31(17) :2877–2878.
- [133] Gaulke, C. A. and Sharpston, T. J. (2018). The influence of ethnicity and geography on human gut microbiome composition. *Nature Medicine*, 24(10) :1495–1496.
- [134] Ge, H., Sun, L., and Yu, J. (2017). Fast batch searching for protein homology based on compression and clustering. *BMC bioinformatics*, 18(1) :508.
- [135] Geleijnse, J. M., Vermeer, C., Grobbee, D. E., Schurgers, L. J., Knapen, M. H. J., van der Meer, I. M., Hofman, A., and Witteman, J. C. M. (2004). Dietary Intake of Menaquinone Is Associated with a Reduced Risk of Coronary Heart Disease : The Rotterdam Study. *The Journal of Nutrition*, 134(11) :3100–3105.
- [136] Ghemawat, S. and Dean, J. (2011). LevelDB.

- [137] Gibson, M. K., Forsberg, K. J., and Dantas, G. (2015). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *The ISME Journal*, 9(1) :207–216.
- [138] Gire, S. K., Goba, A., Andersen, K. G., Sealfon, R. S. G., Park, D. J., Kanneh, L., Jalloh, S., Momoh, M., Fullah, M., Dudas, G., Wohl, S., Moses, L. M., Yozwiak, N. L., Winnicki, S., Matranga, C. B., Malboeuf, C. M., Qu, J., Gladden, A. D., Schaffner, S. F., Yang, X., Jiang, P.-P., Nekoui, M., Colubri, A., Coomber, M. R., Fonnies, M., Moigboi, A., Gbakie, M., Kamara, F. K., Tucker, V., Konuwa, E., Saffa, S., Sellu, J., Jalloh, A. A., Kovoma, A., Koninga, J., Mustapha, I., Kargbo, K., Foday, M., Yillah, M., Kanneh, F., Robert, W., Massally, J. L. B., Chapman, S. B., Bochicchio, J., Murphy, C., Nusbaum, C., Young, S., Birren, B. W., Grant, D. S., Scheffelin, J. S., Lander, E. S., Happi, C., Gevaio, S. M., Gnirke, A., Rambaut, A., Garry, R. F., Khan, S. H., and Sabeti, P. C. (2014). Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202) :1369–1372.
- [139] Glaeser, S. P. and Kämpfer, P. (2015). Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Systematic and applied microbiology*, 38(4) :237–45.
- [140] Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome Datasets Are Compositional : And This Is Not Optional. *Frontiers in Microbiology*, 8.
- [141] Goh, V. H. H., Tain, C. F., Tong, T. Y. Y., Mok, H. P. P., and Wong, M. T. (2004). Are BMI and other anthropometric measures appropriate as indices for obesity? A study in an Asian population. *Journal of Lipid Research*, 45(10) :1892–1898.
- [142] Gominak, S. C. (2016). Vitamin D deficiency changes the intestinal microbiome reducing B vitamin production in the gut. The resulting lack of pantothenic acid adversely affects the immune system, producing a “pro-inflammatory” state associated with atherosclerosis and autoimmun. *Medical Hypotheses*, 94 :103–107.
- [143] Gontarz, P. M., Berger, J., and Wong, C. F. (2013). SRmapper : a fast and sensitive genome-hashing alignment tool. *Bioinformatics (Oxford, England)*, 29(3) :316–21.
- [144] Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age : ten years of next-generation sequencing technologies. *Nature reviews. Genetics*, 17(6) :333–51.
- [145] Google (2008). Protocol Buffers.
- [146] Goto, N., Prins, P., Nakao, M., Bonnal, R., Aerts, J., and Katayama, T. (2010). BioRuby : bioinformatics software for the Ruby programming language. *Bioinformatics*, 26(20) :2617–2619.

- [147] Gotoh, O. (1996). Significant Improvement in Accuracy of Multiple Protein Sequence Alignments by Iterative Refinement as Assessed by Reference to Structural Alignments. *Journal of Molecular Biology*, 264(4) :823–838.
- [148] Gu, S., Fang, L., and Xu, X. (2013). Using SOAPaligner for Short Reads Alignment. *Current protocols in bioinformatics*, 44 :11.11.1–17.
- [149] Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies : Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3) :307–321.
- [150] Guindon, S. and Gascuel, O. (2002). Efficient biased estimation of evolutionary distances when substitution rates vary across sites. *Molecular Biology and Evolution*, 19(4) :534–543.
- [151] Guindon, S. and Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, 52(5) :696–704.
- [152] Guo, Z., Zhang, J., Wang, Z., Ang, K. Y., Huang, S., Hou, Q., Su, X., Qiao, J., Zheng, Y., Wang, L., Koh, E., Danliang, H., Xu, J., Lee, Y. K., and Zhang, H. (2016). Intestinal Microbiota Distinguish Gout Patients from Healthy Humans. *Scientific Reports*, 6(1) :20602.
- [153] Gupta, S. K., Padmanabhan, B. R., Diene, S. M., Lopez-Rojas, R., Kempf, M., Landraud, L., and Rolain, J.-M. (2014). ARG-ANNOT, a New Bioinformatic Tool To Discover Antibiotic Resistance Genes in Bacterial Genomes. *Antimicrobial Agents and Chemotherapy*, 58(1) :212–220.
- [154] Gupta, V. K., Paul, S., and Dutta, C. (2017). Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity. *Frontiers in Microbiology*, 8.
- [155] Hach, F., Hormozdiari, F., Alkan, C., Hormozdiari, F., Birol, I., Eichler, E. E., and Sahinalp, S. C. (2010). mrsFAST : a cache-oblivious algorithm for short-read mapping. *Nature methods*, 7(8) :576–7.
- [156] Harrington, C. T., Lin, E. I., Olson, M. T., and Eshleman, J. R. (2013). Fundamentals of pyrosequencing. *Archives of pathology & laboratory medicine*, 137(9) :1296–303.
- [157] Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., DiMeo, J., Efcavitch, J. W., Giladi, E., Gill, J., Healy, J., Jarosz, M., Lapen, D., Moulton, K., Quake, S. R., Steinmann, K., Thayer, E., Tyurina, A., Ward, R., Weiss, H., and Xie, Z. (2008). Single-Molecule DNA Sequencing of a Viral Genome. *Science*, 320(5872) :106–109.
- [158] Hastie, T., Rosset, S., Zhu, J., and Zou, H. (2009). Multi-class AdaBoost. *Statistics and Its Interface*, 2(3) :349–360.

- [159] Haubold, B. (2014). Alignment-free phylogenetics and population genetics. *Briefings in bioinformatics*, 15(3) :407–18.
- [160] Hazen, T. H., Pan, L., Gu, J.-D., and Sobecky, P. A. (2010). The contribution of mobile genetic elements to the evolution and ecology of Vibrios. *FEMS microbiology ecology*, 74(3) :485–99.
- [161] Hearst, M. A. (1998). Support Vector Machines. *IEEE Intelligent Systems*, 13(4) :18–28.
- [162] Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89(22) :10915–9.
- [163] Herrmann, K. M. (1995). The Shikimate Pathway : Early Steps in the Biosynthesis of Aromatic Compounds. *The Plant Cell*, pages 907–919.
- [164] Hess, E. L. (1970). Origins of molecular biology. *Science (New York, N.Y.)*, 168(3932) :664–9.
- [165] Hewitt, C. E. (1977). Viewing control structures as patterns of message passing. *Artificial intelligence*, 8(3) :323–364.
- [166] Hijova, E. and Chmelarova, A. (2007). Short chain fatty acids and colonic health. *Bratislavske lekarske listy*, 108(8) :354–8.
- [167] Hilty, M., Wüthrich, D., Salter, S. J., Engel, H., Campbell, S., Sá-Leão, R., de Lencastre, H., Hermans, P., Sadowy, E., Turner, P., Chewapreecha, C., Diggle, M., Pluschke, G., McGee, L., Eser, Ö. K., Low, D. E., Smith-Vaughan, H., Endimiani, A., Küffer, M., Dupasquier, M., Beaudoin, E., Weber, J., Bruggmann, R., Hanage, W. P., Parkhill, J., Hathaway, L. J., Mühlemann, K., and Bentley, S. D. (2014). Global phylogenomic analysis of nonencapsulated *Streptococcus pneumoniae* reveals a deep-branching classic lineage that is distinct from multiple sporadic lineages. *Genome biology and evolution*, 6(12) :3281–94.
- [168] Ho, T. K. (1995). Random Decision Forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR '95, page 278, USA. IEEE Computer Society.
- [169] Homer, N., Merriman, B., and Nelson, S. F. (2009). BFAST : An Alignment Tool for Large Scale Genome Resequencing. *PLoS ONE*, 4(11) :e7767.
- [170] Hrdina, J., Banning, A., Kipp, A., Loh, G., Blaut, M., and Brigelius-Flohé, R. (2009). The gastrointestinal microbiota affects the selenium status and selenoprotein expression in mice. *The Journal of Nutritional Biochemistry*, 20(8) :638–648.

- [171] Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3 : Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular biology and evolution*, 33(6) :1635–8.
- [172] Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., and Bork, P. (2019). eggNOG 5.0 : a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1) :D309–D314.
- [173] Hurt, R. T., Kulisek, C., Buchanan, L. A., and McClave, S. A. (2010). The obesity epidemic : challenges, health initiatives, and implications for gastroenterologists. *Gastroenterology & hepatology*, 6(12) :780.
- [174] Husi, H. (2019). *Computational Biology*. Codon Publications.
- [175] Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., and Tappu, R. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLOS Computational Biology*, 12(6) :e1004957.
- [176] Integrative HMP (iHMP) Research Network Consortium (2014). The integrative human microbiome project : Dynamic analysis of microbiome-host omics profiles during periods of human health and disease corresponding author. *Cell Host and Microbe*, 16(3) :276–289.
- [177] International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(7063) :1299–320.
- [178] Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, 44(2) :226–232.
- [179] Jackson, D. A., Symons, R. H., and Berg, P. (1972). Biochemical method for inserting new genetic information into DNA of Simian Virus 40 : circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 69(10) :2904–9.
- [180] Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*, 9(1) :5114.
- [181] Jelic, K., Luzio, S. D., Dunseath, G., Colding-Jorgensen, M., and Owens, D. R. (2007). A Cross-Sectional Analysis of NEFA Levels Following a Standard Mixed Meal in a Population of Persons with Newly Diagnosed Type 2 Diabetes Mellitus Across a Spectrum of Glycemic Control. *Diabetes*, 56.

- [182] Jensen, L. J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T., and Bork, P. (2007). eggNOG : automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research*, 36(Database) :D250–D254.
- [183] Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., Lago, B. A., Dave, B. M., Pereira, S., Sharma, A. N., Doshi, S., Courtot, M., Lo, R., Williams, L. E., Frye, J. G., Elsayegh, T., Sardar, D., Westman, E. L., Pawlowski, A. C., Johnson, T. A., Brinkman, F. S., Wright, G. D., and McArthur, A. G. (2017). CARD 2017 : expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, 45(D1) :D566–D573.
- [184] Jiang, H. and Wong, W. H. (2008). SeqMap : mapping massive amount of oligonucleotides to the genome. *Bioinformatics (Oxford, England)*, 24(20) :2395–6.
- [185] Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., and Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, 10(1) :5029.
- [186] Johnson, L. S., Eddy, S. R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, 11(1) :431.
- [187] Jolley, K. A., Bray, J. E., and Maiden, M. C. J. (2018). Open-access bacterial population genomics : BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Research*, 3 :124.
- [188] Jones, E., Oliphant, T., Peterson, P., and Others (2001). SciPy : Open source scientific tools for Python.
- [189] Kahn, S. E., Hull, R. L., and Utzschneider, K. M. (2006). Mechanisms linking obesity to insulin resistance and type 2 diabetes. *Nature*, 444(7121) :840–846.
- [190] Kanehisa, M. (2000). KEGG : Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1) :27–30.
- [191] Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG : new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1) :D353–D361.
- [192] Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1) :D457–D462.
- [193] Kapli, P., Yang, Z., and Telford, M. J. (2020). Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*.

- [194] Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C. J., Fagerberg, B., Nielsen, J., and Bäckhed, F. (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*, 498(7452) :99–103.
- [195] Karp, P. D. (2002). The MetaCyc Database. *Nucleic Acids Research*, 30(1) :59–61.
- [196] Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., Keseler, I. M., Krummenacker, M., Midford, P. E., Ong, Q., Ong, W. K., Paley, S. M., and Subhraveti, P. (2019). The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics*, 20(4) :1085–1093.
- [197] Karp, P. D., Ong, W. K., Paley, S., Billington, R., Caspi, R., Fulcher, C., Kothari, A., Krummenacker, M., Latendresse, M., Midford, P. E., Subhraveti, P., Gama-Castro, S., Muñoz-Rascado, L., Bonavides-Martinez, C., Santos-Zavaleta, A., Mackie, A., Collado-Vides, J., Keseler, I. M., and Paulsen, I. (2018). The EcoCyc Database. *EcoSal Plus*, 8(1).
- [198] Katoh, K. (2002). MAFFT : a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14) :3059–3066.
- [199] Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7 : Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4) :772–780.
- [200] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm : A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pages 3146–3154.
- [201] Keegan, K. P., Glass, E. M., and Meyer, F. (2016). MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. In *Microbial Environmental Genomics (MEG)*, pages 207–233. Humana Press.
- [202] Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome research*, 12(4) :656–64.
- [203] Kho, Z. Y. and Lal, S. K. (2018). The Human Gut Microbiome – A Potential Controller of Wellness and Disease. *Frontiers in Microbiology*, 9.
- [204] Kim, J., Kim, M. S., Koh, A. Y., Xie, Y., and Zhan, X. (2016). FMAP : Functional Mapping and Analysis Pipeline for metagenomics and metatranscriptomics studies. *BMC Bioinformatics*, 17(1) :420.
- [205] Kim, K. N., Yao, Y., and Ju, S. Y. (2019). Short Chain Fatty Acids and Fecal Microbiota Abundance in Humans with Obesity : A Systematic Review and Meta-Analysis. *Nutrients*, 11(10) :2512.

- [206] Kitahara, K. and Miyazaki, K. (2013). Revisiting bacterial phylogeny. *Mobile Genetic Elements*, 3(1) :e24210.
- [207] Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciulek, T., McCall, L.-I., McDonald, D., Melnik, A. V., Morton, J. T., Navas, J., Quinn, R. A., Sanders, J. G., Swafford, A. D., Thompson, L. R., Tripathi, A., Xu, Z. Z., Zaneveld, J. R., Zhu, Q., Caporaso, J. G., and Dorrestein, P. C. (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology*, 16(7) :410–422.
- [208] Konstantinidis, K. T., Ramette, A., and Tiedje, J. M. (2006). Toward a more robust assessment of intraspecies diversity, using fewer genetic markers. *Applied and Environmental Microbiology*, 72(11) :7286–7293.
- [209] Konstantinidis, K. T. and Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences*, 102(7) :2567–2572.
- [210] Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu : scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation. *Genome Research*, 27(5) :722–736.
- [211] Kos, V. N., Déraspe, M., McLaughlin, R. E., Whiteaker, J. D., Roy, P. H., Alm, R. A., Corbeil, J., and Gardner, H. (2015). The resistome of *Pseudomonas aeruginosa* in relationship to phenotypic susceptibility. *Antimicrobial Agents and Chemotherapy*, 59(1) :427–436.
- [212] Kothari, C., Osseni, M. A., Agbo, L., Ouellette, G., Déraspe, M., Laviolette, F., Corbeil, J., Lambert, J.-P., Diorio, C., and Durocher, F. (2020). Machine learning analysis identifies genes differentiating triple negative breast cancers. *Scientific Reports*, 10(1) :10464.
- [213] Kuhner, M. K. and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11(3) :459–468.
- [214] Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome biology*, 5(2) :R12.
- [215] Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology*, 11(5) :e1004226.
- [216] Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., Poudel, S., and Ussery, D. W. (2015). Insights from 20 years of bacterial genome sequencing. *Functional & integrative genomics*, 15(2) :141–61.

- [217] Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepille, D. E., Vega Thurber, R. L., Knight, R., Beiko, R. G., and Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, 31(9) :814–821.
- [218] Langmead, B. (2010). Aligning short sequencing reads with Bowtie. *Current protocols in bioinformatics*, Chapter 11 :Unit 11.7.
- [219] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4) :357–9.
- [220] Larkin, M., Blackshields, G., Brown, N., Chenna, R., McGettigan, P., McWilliam, H., Valentin, F., Wallace, I., Wilm, A., Lopez, R., Thompson, J., Gibson, T., and Higgins, D. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21) :2947–2948.
- [221] Larsson, P., Elfsmark, D., Svensson, K., Wikström, P., Forsman, M., Brettin, T., Keim, P., and Johansson, A. (2009). Molecular evolutionary consequences of niche restriction in *Francisella tularensis*, a facultative intracellular pathogen. *PLoS pathogens*, 5(6) :e1000472.
- [222] Lassalle, F., Périan, S., Bataillon, T., Nesme, X., Duret, L., and Daubin, V. (2015). GC-Content evolution in bacterial genomes : the biased gene conversion hypothesis expands. *PLoS genetics*, 11(2) :e1004941.
- [223] Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J.-M., Kennedy, S., Leonard, P., Li, J., Burgdorf, K., Grarup, N., Jørgensen, T., Brandslund, I., Nielsen, H. B., Juncker, A. S., Bertalan, M., Levenez, F., Pons, N., Rasmussen, S., Sunagawa, S., Tap, J., Tims, S., Zoetendal, E. G., Brunak, S., Clément, K., Doré, J., Kleerebezem, M., Kristiansen, K., Renault, P., Sicheritz-Ponten, T., de Vos, W. M., Zucker, J.-D., Raes, J., Hansen, T., consortium, M., Bork, P., Wang, J., Ehrlich, S. D., and Pedersen, O. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500(7464) :541–6.
- [224] Lee, J., Lee, J.-Y., Lee, J.-H., Jung, S.-M., Suh, Y. S., Koh, J.-H., Kwok, S.-K., Ju, J. H., Park, K.-S., and Park, S.-H. (2015). Visceral fat obesity is highly associated with primary gout in a metabolically obese but normal weighted population : a case control study. *Arthritis Research & Therapy*, 17(1) :79.
- [225] Lee, W.-P., Stromberg, M. P., Ward, A., Stewart, C., Garrison, E. P., and Marth, G. T. (2014). MOSAIK : a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PloS one*, 9(3) :e90581.
- [226] Leonel, A. J. and Alvarez-Leite, J. I. (2012). Butyrate : Implications for Intestinal Function. *Current Opinion in Clinical Nutrition and Metabolic Care*, 15(5) :474–479.

- [227] Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., and Webb, W. W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. *Science (New York, N.Y.)*, 299(5607) :682–6.
- [228] Ley, R. E., Backhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., and Gordon, J. I. (2005). Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences*, 102(31) :11070–11075.
- [229] Ley, R. E., Turnbaugh, P. J., Klein, S., and Gordon, J. I. (2006). Human gut microbes associated with obesity. *Nature*, 444(7122) :1022–1023.
- [230] Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT : an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31(10) :1674–1676.
- [231] Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., and Lam, T.-W. (2016). MEGAHIT v1.0 : A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, 102 :3–11.
- [232] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14) :1754–60.
- [233] Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5) :589–95.
- [234] Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11) :1851–1858.
- [235] Li, Y. and XifengYan (2015). MSPKmerCounter : A Fast and Memory Efficient Approach for K-mer Counting. *Cs.Ucsb.Edu*, pages 1–7.
- [236] Litvak, Y., Byndloss, M. X., and Bäumlner, A. J. (2018). Colonocyte metabolism shapes the gut microbiota. *Science (New York, N.Y.)*, 362(6418).
- [237] Liu, B. and Pop, M. (2009). ARDB–Antibiotic Resistance Genes Database. *Nucleic Acids Research*, 37(Database) :D443–D447.
- [238] Liu, B., Zheng, D., Jin, Q., Chen, L., and Yang, J. (2019). VFDB 2019 : a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Research*, 47(D1) :D687–D692.
- [239] Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology*, 2012 :251364.

- [240] Liu, M., Li, X., Xie, Y., Bi, D., Sun, J., Li, J., Tai, C., Deng, Z., and Ou, H.-Y. (2019). ICEberg 2.0 : an updated database of bacterial integrative and conjugative elements. *Nucleic Acids Research*, 47(D1) :D660–D665.
- [241] Loureiro, A., Torgo, L., and Soares, C. (2004). Outlier detection using clustering methods : a data cleaning application. In *Proceedings of KDNet Symposium on Knowledge-based systems for the Public Sector*.
- [242] Lowe, T. M. and Eddy, S. R. (1997). tRNAscan-SE : A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Research*, 25(5) :955–964.
- [243] Lozupone, C. A., Hamady, M., Kelley, S. T., and Knight, R. (2007). Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. *Applied and Environmental Microbiology*, 73(5) :1576–1585.
- [244] Lu, H., Giordano, F., and Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, proteomics & bioinformatics*, 14(5) :265–279.
- [245] Lu, J., Breitwieser, F. P., Thielen, P., and Salzberg, S. L. (2017). Bracken : estimating species abundance in metagenomics data. *PeerJ Computer Science*, 3 :e104.
- [246] Lu, L., Pillai, T. S., Arpaci-Dusseau, A. C., and Arpaci-Dusseau, R. H. (2016). WisKey : Separating Keys from Values in SSD-conscious Storage. In *14th USENIX Conference on File and Storage Technologies (FAST 16)*, pages 133–148, Santa Clara, CA. USENIX Association.
- [247] Lunter, G. and Goodson, M. (2011). Stampy : A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, 21(6) :936–939.
- [248] Magnúsdóttir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., Greenhalgh, K., Jäger, C., Baginska, J., Wilmes, P., Fleming, R. M. T., and Thiele, I. (2017). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nature Biotechnology*, 35(1) :81–89.
- [249] Magoc, T., Pabinger, S., Canzar, S., Liu, X., Su, Q., Puiu, D., Tallon, L. J., and Salzberg, S. L. (2013). GAGE-B : an evaluation of genome assemblers for bacterial organisms. *Bioinformatics*, 29(14) :1718–1725.
- [250] Malla, M. A., Dubey, A., Kumar, A., Yadav, S., Hashem, A., and Abd-Allah, E. F. (2019). Exploring the Human Microbiome : The Potential Future Role of Next-Generation Sequencing in Disease Diagnosis and Treatment. *Frontiers in Immunology*, 9.
- [251] Marçais, G. and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6) :764–770.

- [252] Marchand, M. and Shawe-Taylor, J. (2002). The set covering machine. *Journal of Machine Learning Research*, 3(Dec) :723–746.
- [253] Mardinoglu, A., Shoaie, S., Bergentall, M., Ghaffari, P., Zhang, C., Larsson, E., Bäckhed, F., and Nielsen, J. (2015). The gut microbiota modulates host amino acid and glutathione metabolism in mice. *Molecular Systems Biology*, 11(10) :834.
- [254] Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057) :376–80.
- [255] Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P., Huntemann, M., Anderson, I., Mavromatis, K., Ivanova, N. N., and Kyrpides, N. C. (2012). IMG : the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research*, 40(D1) :D115–D122.
- [256] Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1) :10.
- [257] Materon, I. C., Queenan, A. M., Koehler, T. M., Bush, K., and Palzkill, T. (2003). Biochemical characterization of beta-lactamases Bla1 and Bla2 from *Bacillus anthracis*. *Antimicrobial agents and chemotherapy*, 47(6) :2040–2.
- [258] McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., Bhullar, K., Canova, M. J., De Pascale, G., Ejim, L., Kalan, L., King, A. M., Koteva, K., Morar, M., Mulvey, M. R., O’Brien, J. S., Pawlowski, A. C., Piddock, L. J. V., Spanogiannopoulos, P., Sutherland, A. D., Tang, I., Taylor, P. L., Thaker, M., Wang, W., Yan, M., Yu, T., and Wright, G. D. (2013). The Comprehensive Antibiotic Resistance Database. *Antimicrobial Agents and Chemotherapy*, 57(7) :3348–3357.
- [259] Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., de Bruijn, I., Chooi, Y. H., Claesen, J., Coates, R. C., Cruz-Morales, P., Duddela, S., Düsterhus, S., Edwards, D. J., Fewer, D. P., Garg, N., Geiger, C., Gomez-Escribano, J. P., Greule, A., Hadjithomas, M., Haines, A. S., Helfrich, E. J. N., Hillwig, M. L., Ishida, K., Jones, A. C., Jones, C. S., Jungmann, K., Kegler, C., Kim, H. U., Kötter, P., Krug, D., Masschelein, J., Melnik, A. V., Mantovani, S. M., Monroe, E. A., Moore, M., Moss, N.,

- Nützmann, H.-W., Pan, G., Pati, A., Petras, D., Reen, F. J., Rosconi, F., Rui, Z., Tian, Z., Tobias, N. J., Tsunematsu, Y., Wiemann, P., Wyckoff, E., Yan, X., Yim, G., Yu, F., Xie, Y., Aigle, B., Apel, A. K., Balibar, C. J., Balskus, E. P., Barona-Gómez, F., Bechthold, A., Bode, H. B., Borriss, R., Brady, S. F., Brakhage, A. A., Caffrey, P., Cheng, Y.-Q., Clardy, J., Cox, R. J., De Mot, R., Donadio, S., Donia, M. S., van der Donk, W. A., Dorrestein, P. C., Doyle, S., Driessen, A. J. M., Ehling-Schulz, M., Entian, K.-D., Fischbach, M. A., Gerwick, L., Gerwick, W. H., Gross, H., Gust, B., Hertweck, C., Höfte, M., Jensen, S. E., Ju, J., Katz, L., Kaysser, L., Klassen, J. L., Keller, N. P., Kormanec, J., Kuipers, O. P., Kuzuyama, T., Kyrpides, N. C., Kwon, H.-J., Lautru, S., Lavigne, R., Lee, C. Y., Linquan, B., Liu, X., Liu, W., Luzhetskyy, A., Mahmud, T., Mast, Y., Méndez, C., Metsä-Ketelä, M., Micklefield, J., Mitchell, D. A., Moore, B. S., Moreira, L. M., Müller, R., Neilan, B. A., Nett, M., Nielsen, J., O’Gara, F., Oikawa, H., Osbourn, A., Osburne, M. S., Ostash, B., Payne, S. M., Pernodet, J.-L., Petricek, M., Piel, J., Ploux, O., Raaijmakers, J. M., Salas, J. A., Schmitt, E. K., Scott, B., Seipke, R. F., Shen, B., Sherman, D. H., Sivonen, K., Smanski, M. J., Sosio, M., Stegmann, E., Süßmuth, R. D., Tahlan, K., Thomas, C. M., Tang, Y., Truman, A. W., Viaud, M., Walton, J. D., Walsh, C. T., Weber, T., van Wezel, G. P., Wilkinson, B., Willey, J. M., Wohlleben, W., Wright, G. D., Ziemert, N., Zhang, C., Zotchev, S. B., Breitling, R., Takano, E., and Glöckner, F. O. (2015). Minimum Information about a Biosynthetic Gene cluster. *Nature chemical biology*, 11(9) :625–31.
- [260] Medini, D., Donati, C., Tettelin, H., Massignani, V., and Rappuoli, R. (2005). The microbial pan-genome. *Current opinion in genetics & development*, 15(6) :589–94.
- [261] Melsted, P. and Pritchard, J. K. (2011). Efficient counting of k-mers in DNA sequences using a bloom filter. *BMC bioinformatics*, 12(1) :333.
- [262] Metcalf, J. A., Funkhouser-Jones, L. J., Brileya, K., Reysenbach, A.-L., and Bordenstein, S. R. (2014). Antibacterial gene transfer across the tree of life. *eLife*, 3.
- [263] Milani, C., Ticinesi, A., Gerritsen, J., Nouvenne, A., Lugli, G. A., Mancabelli, L., Turroni, F., Duranti, S., Mangifesta, M., Viappiani, A., Ferrario, C., Maggio, M., Lauretani, F., De Vos, W., van Sinderen, D., Meschi, T., and Ventura, M. (2016). Gut microbiota composition and *Clostridium difficile* infection in hospitalized elderly individuals : a metagenomic study. *Scientific Reports*, 6(1) :25945.
- [264] Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., and Punta, M. (2013). Challenges in homology search : HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*, 41(12) :e121–e121.
- [265] Mitchell, A. L., Almeida, A., Beracochea, M., Boland, M., Burgin, J., Cochrane, G., Crusoe, M. R., Kale, V., Potter, S. C., Richardson, L. J., Sakharova, E., Scheremetjew, M., Korobeynikov, A., Shlemov, A., Kunyavskaya, O., Lapidus, A., and Finn, R. D. (2019). MGnify : the microbiome analysis resource in 2020. *Nucleic Acids Research*.

- [266] Mithieux, G. (2009). A novel function of intestinal gluconeogenesis : Central signaling in glucose and energy homeostasis. *Nutrition*, 25(9) :881–884.
- [267] Mooers and Holmes (2000). The evolution of base composition and phylogenetic inference. *Trends in ecology & evolution*, 15(9) :365–369.
- [268] Moura, A., Soares, M., Pereira, C., Leitao, N., Henriques, I., and Correia, A. (2009). INTEGRALL : a database and search engine for integrons, integrases and gene cassettes. *Bioinformatics*, 25(8) :1096–1098.
- [269] Mullis, K. B. and Faloona, F. A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods in enzymology*, 155 :335–50.
- [270] Narayan, V., Kudva, A. K., and Prabhu, K. S. (2015). Reduction of Tetrathionate by Mammalian Thioredoxin Reductase. *Biochemistry*, 54(33) :5121–5124.
- [271] Nasser, W., Beres, S. B., Olsen, R. J., Dean, M. A., Rice, K. A., Long, S. W., Kristinsson, K. G., Gottfredsson, M., Vuopio, J., Raisanen, K., Caugant, D. A., Steinbakk, M., Low, D. E., McGeer, A., Darenberg, J., Henriques-Normark, B., Van Beneden, C. A., Hoffmann, S., and Musser, J. M. (2014). Evolutionary pathway to increased virulence and epidemic group A Streptococcus disease derived from 3,615 genome sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 111(17) :E1768–76.
- [272] Nayfach, S., Bradley, P. H., Wyman, S. K., Laurent, T. J., Williams, A., Eisen, J. A., Polard, K. S., and Sharpton, T. J. (2015). Automated and Accurate Estimation of Gene Family Abundance from Shotgun Metagenomes. *PLOS Computational Biology*, 11(11) :e1004573.
- [273] Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3) :443–453.
- [274] Neis, E., Dejong, C., and Rensen, S. (2015). The Role of Microbial Amino Acid Metabolism in Host Metabolism. *Nutrients*, 7(4) :2930–2946.
- [275] Ni, J., Wu, G. D., Albenberg, L., and Tomov, V. T. (2017). Gut microbiota and IBD : Causation or correlation? *Nature Reviews Gastroenterology and Hepatology*, 14(10) :573–584.
- [276] Niehaus, K. E., Walker, T. M., Crook, D. W., Peto, T. E. A., and Clifton, D. A. (2014). Machine learning for the prediction of antibacterial susceptibility in Mycobacterium tuberculosis. In *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 618–621. IEEE.
- [277] Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D. R., Gautier, L., Pedersen, A. G., Le Chatelier, E., Pelletier, E., Bonde, I., Nielsen,

- T., Manichanh, C., Arumugam, M., Batto, J.-M., Quintanilha Dos Santos, M. B., Blom, N., Borruel, N., Burgdorf, K. S., Boumezbeur, F., Casellas, F., Doré, J., Dworzynski, P., Guarner, F., Hansen, T., Hildebrand, F., Kaas, R. S., Kennedy, S., Kristiansen, K., Kultima, J. R., Léonard, P., Levenez, F., Lund, O., Moumen, B., Le Paslier, D., Pons, N., Pedersen, O., Prifti, E., Qin, J., Raes, J., Sørensen, S., Tap, J., Tims, S., Ussery, D. W., Yamada, T., MetaHIT Consortium, Renault, P., Sicheritz-Ponten, T., Bork, P., Wang, J., Brunak, S., Ehrlich, S. D., and MetaHIT Consortium (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature biotechnology*, 32(8) :822–8.
- [278] NIH HMP Working Group, Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., Bonazzi, V., McEwen, J. E., Wetterstrand, K. A., Deal, C., Baker, C. C., Di Francesco, V., Howcroft, T. K., Karp, R. W., Lunsford, R. D., Wellington, C. R., Belachew, T., Wright, M., Giblin, C., David, H., Mills, M., Salomon, R., Mullins, C., Akolkar, B., Begg, L., Davis, C., Grandison, L., Humble, M., Khalsa, J., Little, A. R., Peavy, H., Pontzer, C., Portnoy, M., Sayre, M. H., Starke-Reed, P., Zakhari, S., Read, J., Watson, B., and Guyer, M. (2009). The NIH Human Microbiome Project. *Genome research*, 19(12) :2317–23.
- [279] Notredame, C. (1996). SAGA : sequence alignment by genetic algorithm. *Nucleic Acids Research*, 24(8) :1515–1524.
- [280] Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee : a novel method for fast and accurate multiple sequence alignment 1 Edited by J. Thornton. *Journal of Molecular Biology*, 302(1) :205–217.
- [281] Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. A. (2017). metaSPAdes : a new versatile metagenomic assembler. *Genome Research*, 27(5) :824–834.
- [282] Nuttall, F. Q. (2015). Body Mass Index. *Nutrition Today*, 50(3) :117–128.
- [283] Nyren, P., Pettersson, B., and Uhlen, M. (1993). Solid Phase DNA Minisequencing by an Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay. *Analytical Biochemistry*, 208(1) :171–175.
- [284] Obregon-Tito, A. J., Tito, R. Y., Metcalf, J., Sankaranarayanan, K., Clemente, J. C., Ursell, L. K., Zech Xu, Z., Van Treuren, W., Knight, R., Gaffney, P. M., Spicer, P., Lawson, P., Marin-Reyes, L., Trujillo-Villarroel, O., Foster, M., Guija-Poma, E., Troncoso-Corzo, L., Warinner, C., Ozga, A. T., and Lewis, C. M. (2015). Subsistence strategies in traditional societies distinguish gut microbiomes. *Nature Communications*, 6(1) :6505.
- [285] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG : Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27(1) :29–34.

- [286] Ohler, U., Liao, G.-c., Niemann, H., and Rubin, G. M. (2002). Computational analysis of core promoters in the *Drosophila* genome. *Genome biology*, 3(12) :research0087—1.
- [287] O’Leary, N. A., Wright, M. W., Brister, J. R., Ciufu, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O’Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI : current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1) :D733–D745.
- [288] O’Leary, N. A., Wright, M. W., Brister, J. R., Ciufu, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O’Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI : current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1) :D733–45.
- [289] Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., and Phillippy, A. M. (2016). Mash : Fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17(1) :132.
- [290] O’Neil, P., Cheng, E., Gawlick, D., and O’Neil, E. (1996). The log-structured merge-tree (LSM-tree). *Acta Informatica*, 33(4) :351–385.
- [291] Onyemaechi, N., Anyanwu, G., Obikili, E., Onwuasoigwe, O., and Nwankwo, O. (2016). Impact of overweight and obesity on the musculoskeletal system using lumbosacral angles. *Patient Preference and Adherence*, page 291.
- [292] Ounit, R. and Lonardi, S. (2016). Higher classification sensitivity of short metagenomic reads with CLARK- S. *Bioinformatics*, 32(24) :3823–3825.
- [293] Paley, E. L. (2019). Diet-Related Metabolic Perturbations of Gut Microbial Shikimate Pathway-Tryptamine-tRNA Aminoacylation-Protein Synthesis in Human Health and Disease. *International Journal of Tryptophan Research*, 12 :117864691983455.

- [294] Paradis, E., Claude, J., and Strimmer, K. (2004). APE : Analyses of Phylogenetics and Evolution in R language. *Bioinformatics (Oxford, England)*, 20(2) :289–90.
- [295] Pärnänen, K., Karkman, A., Tamminen, M., Lyra, C., Hultman, J., Paulin, L., and Virta, M. (2016). Evaluating the mobility potential of antibiotic resistance genes in environmental resistomes without metagenomics. *Scientific reports*, 6 :35790.
- [296] Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., Beghini, F., Malik, F., Ramos, M., Dowd, J. B., Huttenhower, C., Morgan, M., Segata, N., and Waldron, L. (2017). Accessible, curated metagenomic data through ExperimentHub. *Nature Methods*, 14(11) :1023–1024.
- [297] Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine Learning Meta-analysis of Large Metagenomic Datasets : Tools and Biological Insights. *PLOS Computational Biology*, 12(7) :e1004977.
- [298] Patwardhan, A., Ray, S., and Roy, A. (2014). Molecular Markers in Phylogenetic Studies- A Review. *Journal of Phylogenetics & Evolutionary Biology*, 2014.
- [299] Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. John Wiley & Sons.
- [300] Pearson, W. R. (2016). Finding Protein and Nucleotide Similarities with FASTA. *Current protocols in bioinformatics*, 53 :3.9.1–3.9.25.
- [301] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830.
- [302] Pennisi, E. (2008). Evolution. Building the tree of life, genome by genome. *Science (New York, N.Y.)*, 320(5884) :1716–7.
- [303] Pesesky, M. W., Hussain, T., Wallace, M., Patel, S., Andleeb, S., Burnham, C.-A. D., and Dantas, G. (2016). Evaluation of Machine Learning and Rules-Based Approaches for Predicting Antimicrobial Resistance Profiles in Gram-negative Bacilli from Whole Genome Sequence Data. *Frontiers in Microbiology*, 7.
- [304] Pevzner, P. A. and Tang, H. (2001). Fragment assembly with double-barreled data. *Bioinformatics*, 17(Suppl 1) :S225–S233.
- [305] Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17) :9748–9753.

- [306] Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, 98(17) :9748–53.
- [307] Philippe, H. and Douady, C. J. (2003). Horizontal gene transfer and phylogenetics. *Current Opinion in Microbiology*, 6(5) :498–505.
- [308] Pischon, T. and Nimptsch, K. (2016). *Obesity and Cancer*, volume 208 of *Recent Results in Cancer Research*. Springer International Publishing, Cham.
- [309] Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree : Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution*, 26(7) :1641–1650.
- [310] Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2—approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3) :e9490.
- [311] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). CatBoost : unbiased boosting with categorical features. In *Advances in neural information processing systems*, pages 6638–6648.
- [312] Prokop, J. W., May, T., Strong, K., Bilinovich, S. M., Bupp, C., Rajasekaran, S., Worthey, E. A., and Lazar, J. (2018). Genome sequencing in the clinic : the past, present, and future of genomic medicine. *Physiological genomics*, 50(8) :563–579.
- [313] Proneth, B. and Conrad, M. (2019). Ferroptosis and necroinflammation, a yet poorly explored link. *Cell Death & Differentiation*, 26(1) :14–24.
- [314] Qi, J., Wang, B., and Hao, B. I. (2004). Whole Proteome Prokaryote Phylogeny Without Sequence Alignment : A K-String Composition Approach. *Journal of Molecular Evolution*, 58(1) :1–11.
- [315] Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S. D., and Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285) :59–65.
- [316] Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., Wu, P., Dai, Y., Sun, X., Li, Z., Tang, A., Zhong, S., Li, X., Chen, W., Xu, R., Wang, M., Feng,

- Q., Gong, M., Yu, J., Zhang, Y., Zhang, M., Hansen, T., Sanchez, G., Raes, J., Falony, G., Okuda, S., Almeida, M., LeChatelier, E., Renault, P., Pons, N., Batto, J.-M., Zhang, Z., Chen, H., Yang, R., Zheng, W., Li, S., Yang, H., Wang, J., Ehrlich, S. D., Nielsen, R., Pedersen, O., Kristiansen, K., and Wang, J. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418) :55–60.
- [317] Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L., Guo, J., Le Chatelier, E., Yao, J., Wu, L., Zhou, J., Ni, S., Liu, L., Pons, N., Batto, J. M., Kennedy, S. P., Leonard, P., Yuan, C., Ding, W., Chen, Y., Hu, X., Zheng, B., Qian, G., Xu, W., Ehrlich, S. D., Zheng, S., and Li, L. (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature*, 513(7516) :59–64.
- [318] Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., and Gu, Y. (2012). A tale of three next generation sequencing platforms : comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, 13 :341.
- [319] Rasmussen, H. S., Holtug, K., and Mortensen, P. B. (1988). Degradation of amino acids to short-chain fatty acids in humans. An in vitro study. *Scandinavian journal of gastroenterology*, 23(2) :178–82.
- [320] Raymond, F., Boissinot, M., Ouameur, A. A., Déraspe, M., Plante, P.-L., Kpanou, S. R., Bérubé, È., Huletsky, A., Roy, P. H., Ouellette, M., Bergeron, M. G., and Corbeil, J. (2019). Culture-enriched human gut microbiomes reveal core and accessory resistance genes. *Microbiome*, 7(1) :56.
- [321] Raymond, F., Déraspe, M., Boissinot, M., Bergeron, M. G., and Corbeil, J. (2016). Partial recovery of microbiomes after antibiotic treatment. *Gut microbes*, 7(5) :428–34.
- [322] Raymond, F., Ouameur, A. a., Déraspe, M., Iqbal, N., Gingras, H., Dridi, B., Leprohon, P., Plante, P.-L., Giroux, R., Bérubé, È., Frenette, J., Boudreau, D. K., Simard, J.-L., Chabot, I., Domingo, M.-C., Trottier, S., Boissinot, M., Huletsky, A., Roy, P. H., Ouellette, M., Bergeron, M. G., and Corbeil, J. (2015). The initial state of the human gut microbiome determines its reshaping by antibiotics. *The ISME journal*, pages 1–14.
- [323] Reinert, G., Chew, D., Sun, F., and Waterman, M. S. (2009). Alignment-free sequence comparison (I) : statistics and power. *Journal of computational biology : a journal of computational molecular cell biology*, 16(12) :1615–1634.
- [324] Rhee, C. M., Ahmadi, S.-F., and Kalantar-Zadeh, K. (2016). The dual roles of obesity in chronic kidney disease. *Current Opinion in Nephrology and Hypertension*, 25(3) :208–216.
- [325] Rho, M., Tang, H., and Ye, Y. (2010). FragGeneScan : predicting genes in short and error-prone reads. *Nucleic Acids Research*, 38(20) :e191–e191.

- [326] Rizk, G., Lavenier, D., and Chikhi, R. (2013). DSK : K-mer counting with very low memory usage. *Bioinformatics*, 29(5) :652–653.
- [327] Roberts, M., Hayes, W., Hunt, B. R., Mount, S. M., and Yorke, J. A. (2004). Reducing storage requirements for biological sequence comparison. *Bioinformatics*, 20(18) :3363–3369.
- [328] Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53(1-2) :131–147.
- [329] Rodionov, D. A., Gelfand, M. S., Mironov, A. A., and Rakhmaninova, A. B. (2001). Comparative approach to analysis of regulation in complete genomes : multidrug resistance systems in gamma-proteobacteria. *Journal of molecular microbiology and biotechnology*, 3(2) :319–24.
- [330] Roediger, W. E. (1980). Role of anaerobic bacteria in the metabolic welfare of the colonic mucosa in man. *Gut*, 21(9) :793–798.
- [331] Roediger, W. E. W. (1982). Utilization of nutrients by isolated epithelial cells of the rat colon. *Gastroenterology*, 83(2) :424–429.
- [332] Romero, P., Llull, D., García, E., Mitchell, T. J., López, R., and Moscoso, M. (2007). Isolation and characterization of a new plasmid pSpnP1 from a multidrug-resistant clone of *Streptococcus pneumoniae*. *Plasmid*, 58(1) :51–60.
- [333] Ronaghi, M., Uhlén, M., and Nyren, P. (1998). A sequencing method based on real-time pyrophosphate. *Science (New York, N.Y.)*, 281(5375) :363, 365.
- [334] Rosselló-Móra, R. and Amann, R. (2015). Past and future species definitions for Bacteria and Archaea. *Systematic and Applied Microbiology*, 38(4) :209–216.
- [335] Round, J. L. and Mazmanian, S. K. (2009). The gut microbiota shapes intestinal immune responses during health and disease. *Nature Reviews Immunology*, 9(5) :313–323.
- [336] Rumble, S. M., Lacroute, P., Dalca, A. V., Fiume, M., Sidow, A., and Brudno, M. (2009). SHRiMP : accurate mapping of short color-space reads. *PLoS computational biology*, 5(5) :e1000386.
- [337] Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., and Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science (New York, N.Y.)*, 239(4839) :487–91.
- [338] Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., and Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science (New York, N.Y.)*, 230(4732) :1350–4.

- [339] Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12) :5463–5467.
- [340] Sanger, F. and Tuppy, H. (1951). The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochemical Journal*, 49(4) :463–481.
- [341] Sansinenea, E. and Ortiz, A. (2011). Secondary metabolites of soil *Bacillus* spp. *Bio-technology letters*, 33(8) :1523–38.
- [342] Saxena, M. and Yeretssian, G. (2014). NOD-Like Receptors : Master Regulators of Inflammation and Cancer. *Frontiers in Immunology*, 5.
- [343] Sayers, S., Li, L., Ong, E., Deng, S., Fu, G., Lin, Y., Yang, B., Zhang, S., Fa, Z., Zhao, B., Xiang, Z., Li, Y., Zhao, X.-M., Olszewski, M. A., Chen, L., and He, Y. (2019). Victors : a web-based knowledge base of virulence factors in human and animal pathogens. *Nucleic Acids Research*, 47(D1) :D693–D700.
- [344] Schadt, E. E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. *Human molecular genetics*, 19(R2) :R227–40.
- [345] Schirmer, M., Smeeckens, S. P., Vlamakis, H., Jaeger, M., Oosting, M., Franzosa, E. A., ter Horst, R., Jansen, T., Jacobs, L., Bonder, M. J., Kurilshikov, A., Fu, J., Joosten, L. A., Zhernakova, A., Huttenhower, C., Wijmenga, C., Netea, M. G., and Xavier, R. J. (2016). Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. *Cell*, 167(7) :1897.
- [346] Schleifer, K. H. (2009). Classification of Bacteria and Archaea : Past, present and future. *Systematic and Applied Microbiology*, 32(8) :533–542.
- [347] Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009). Annotation error in public databases : Misannotation of molecular function in enzyme superfamilies. *PLoS Computational Biology*, 5(12) :e1000605.
- [348] Schuch, R. and Fischetti, V. A. (2009). The secret life of the anthrax agent *Bacillus anthracis* : bacteriophage-mediated ecological adaptations. *PloS one*, 4(8) :e6532.
- [349] Schwiertz, A., Taras, D., Schäfer, K., Beijer, S., Bos, N. A., Donus, C., and Hardt, P. D. (2010). Microbiota and SCFA in Lean and Overweight Healthy Subjects. *Obesity*, 18(1) :190–195.
- [350] Seabold, S. and Perktold, J. (2010). Statsmodels : Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, page 61. Scipy.

- [351] Seemann, T. (2014). Prokka : rapid prokaryotic genome annotation. *Bioinformatics (Oxford, England)*, 30(14) :2068–9.
- [352] Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I. K., Wang, J.-P. Z., and Widom, J. (2006). A genomic code for nucleosome positioning. *Nature*, 442(7104) :772–778.
- [353] Sender, R., Fuchs, S., and Milo, R. (2016). Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell*, 164(3) :337–340.
- [354] Sender, R., Fuchs, S., and Milo, R. (2016). Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biology*, 14(8) :e1002533.
- [355] Seward, J. (1996). bzip2 and libbzip2. available at <http://www.bzip.org>.
- [356] Shaik, N. A., Hakeem, K. R., Banaganapalli, B., and Elango, R., editors (2019). *Essentials of Bioinformatics, Volume I*. Springer International Publishing, Cham.
- [357] Shanker, A. (2018). *Bioinformatics : Sequences, Structures, Phylogeny*. Springer.
- [358] Shapiro, B. J., Friedman, J., Cordero, O. X., Preheim, S. P., Timberlake, S. C., Szabó, G., Polz, M. F., and Alm, E. J. (2012). Population genomics of early events in the ecological differentiation of bacteria. *Science (New York, N.Y.)*, 336(6077) :48–51.
- [359] Sievers, F. and Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Science*, 27(1) :135–145.
- [360] Siguier, P. (2006). ISfinder : the reference centre for bacterial insertion sequences. *Nucleic Acids Research*, 34(90001) :D32–D36.
- [361] Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., and Birol, I. (2009). ABySS : A parallel assembler for short read sequence data. *Genome Research*, 19(6) :1117–1123.
- [362] Siva, N. (2010). 1000 Genomes Project. *ATLA Alternatives to Laboratory Animals*, 38(6) :445.
- [363] Smith, A. D., Chung, W.-Y., Hodges, E., Kendall, J., Hannon, G., Hicks, J., Xuan, Z., and Zhang, M. Q. (2009). Updates to the RMAP short-read mapping software. *Bioinformatics (Oxford, England)*, 25(21) :2841–2.
- [364] Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, 147(1) :195–7.

- [365] Snitkin, E. S., Zelazny, a. M., Thomas, P. J., Stock, F., Henderson, D. K., Palmore, T. N., and Segre, J. a. (2012). Tracking a Hospital Outbreak of Carbapenem-Resistant *Klebsiella pneumoniae* with Whole-Genome Sequencing. *Science Translational Medicine*, 4(148) :148ra116–148ra116.
- [366] Sokal, R. R. and Rohlf, F. J. (1962). The Comparison of Dendrograms by Objective Methods. *Taxon*, 11(2) :33.
- [367] Song, K., Ren, J., Reinert, G., Deng, M., Waterman, M. S., and Sun, F. (2014). New developments of alignment-free sequence comparison : Measures, statistics and next-generation sequencing. *Briefings in Bioinformatics*, 15(3) :343–353.
- [368] Sozhamannan, S., Chute, M. D., McAfee, F. D., Fouts, D. E., Akmal, A., Galloway, D. R., Mateczun, A., Baillie, L. W., and Read, T. D. (2006). The *Bacillus anthracis* chromosome contains four conserved, excision-proficient, putative prophages. *BMC microbiology*, 6 :34.
- [369] Spellerberg, I. F. and Fedor, P. J. (2003). A tribute to Claude Shannon (1916-2001) and a plea for more rigorous use of species richness, species diversity and the ‘Shannon-Wiener’ Index. *Global Ecology and Biogeography*, 12(3) :177–179.
- [370] Spielman, S. J. and Wilke, C. O. (2015). Pyvolve : A flexible python module for simulating sequences along phylogenies. *PLoS ONE*, 10(9) :e0139047.
- [371] Sprong, R., Schonewille, A., and van der Meer, R. (2010). Dietary cheese whey protein protects rats against mild dextran sulfate sodium–induced colitis : Role of mucin and microbiota. *Journal of Dairy Science*, 93(4) :1364–1371.
- [372] STACKEBRANDT, E. and GOEBEL, B. M. (1994). Taxonomic Note : A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic and Evolutionary Microbiology*, 44(4) :846–849.
- [373] Stamatakis, A. (2006). RAxML-VI-HPC : maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21) :2688–2690.
- [374] Stamatakis, A. (2014). RAxML version 8 : A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9) :1312–1313.
- [375] Sun, Q., Lan, R., Wang, Y., Wang, J., Wang, Y., Li, P., Du, P., and Xu, J. (2013). Isolation and genomic characterization of SflI, a serotype-converting bacteriophage of *Shigella flexneri*. *BMC microbiology*, 13 :39.
- [376] Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. A., Kultima, J. R., Coelho, L. P., Arumugam, M., Tap, J., Nielsen, H. B., Rasmussen, S., Brunak, S.,

- Pedersen, O., Guarner, F., de Vos, W. M., Wang, J., Li, J., Doré, J., Ehrlich, S. D., Stamatidis, A., and Bork, P. (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods*, 10(12) :1196–1199.
- [377] Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., and Consortium, U. (2015). UniRef clusters : a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6) :926–932.
- [378] Tamames, J. and Puente-Sánchez, F. (2019). SqueezeMeta, A Highly Portable, Fully Automatic Metagenomic Analysis Pipeline. *Frontiers in Microbiology*, 9.
- [379] Tang, F., Bossers, A., Harders, F., Lu, C., and Smith, H. (2013). Comparative genomic analysis of twelve *Streptococcus suis* (pro)phages. *Genomics*, 101(6) :336–44.
- [380] Tang, L. (2019). Circular consensus sequencing with long reads. *Nature methods*, 16(10) :958.
- [381] Tatusov, R. L. (2000). The COG database : a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28(1) :33–36.
- [382] Tatusova, T., Ciufu, S., Federhen, S., Fedorov, B., McVeigh, R., O’Neill, K., Tolstoy, I., and Zaslavsky, L. (2015). Update on RefSeq microbial genomes resources. *Nucleic Acids Research*, 43(D1) :D599–D605.
- [383] Tatusova, T., Ciufu, S., Fedorov, B., O’Neill, K., and Tolstoy, I. (2015). RefSeq microbial genomes database : new representation and annotation strategy. *Nucleic acids research*, 43(7) :3872.
- [384] Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M., and Ostell, J. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic acids research*, 44(14) :6614–24.
- [385] The Emerging Risk Factors Collaboration (2011). Separate and combined associations of body-mass index and abdominal adiposity with cardiovascular disease : collaborative analysis of 58 prospective studies. *The Lancet*, 377(9771) :1085–1095.
- [386] The Gene Ontology Consortium (2019). The Gene Ontology Resource : 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1) :D330–D338.
- [387] The Uniprot Consortium (2019). UniProt : a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1) :D506–D515.
- [388] Thompson, J. D., Linard, B., Lecompte, O., and Poch, O. (2011). A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods : Current Challenges and Future Perspectives. *PLoS ONE*, 6(3) :e18093.

- [389] Torres-Fuentes, C., Schellekens, H., Dinan, T. G., and Cryan, J. F. (2017). The microbiota–gut–brain axis in obesity. *The Lancet Gastroenterology & Hepatology*, 2(10) :747–756.
- [390] Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, 12(10) :902–903.
- [391] Tu, Q. and Lin, L. (2016). Gene content dissimilarity for subclassification of highly similar microbial strains. *BMC genomics*, 17 :647.
- [392] Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., Egholm, M., Henrissat, B., Heath, A. C., Knight, R., and Gordon, J. I. (2009). A core gut microbiome in obese and lean twins. *Nature*, 457(7228) :480–484.
- [393] Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The Human Microbiome Project. *Nature*, 449(7164) :804–810.
- [394] Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122) :1027–1031.
- [395] Urban, M., Cuzick, A., Rutherford, K., Irvine, A., Pedro, H., Pant, R., Sadanadan, V., Khamari, L., Billal, S., Mohanty, S., and Hammond-Kosack, K. E. (2017). PHI-base : a new interface and further additions for the multi-species pathogen–host interactions database. *Nucleic Acids Research*, 45(D1) :D604–D610.
- [396] Urban, M., Cuzick, A., Seager, J., Wood, V., Rutherford, K., Venkatesh, S. Y., De Silva, N., Martinez, M. C., Pedro, H., Yates, A. D., Hassani-Pak, K., and Hammond-Kosack, K. E. (2019). PHI-base : the pathogen–host interactions database. *Nucleic Acids Research*.
- [397] Urban, M., Pant, R., Raghunath, A., Irvine, A. G., Pedro, H., and Hammond-Kosack, K. E. (2015). The Pathogen-Host Interactions database (PHI-base) : additions and future developments. *Nucleic Acids Research*, 43(D1) :D645–D655.
- [398] Valitutti, Cucchiara, and Fasano (2019). Celiac Disease and the Microbiome. *Nutrients*, 11(10) :2403.
- [399] van den Nieuwboer, M., van Hemert, S., Claassen, E., and de Vos, W. M. (2016). *Lactobacillus plantarum* WCFS1 and its host interaction : a dozen years after the genome. *Microbial biotechnology*, 9(4) :452–65.
- [400] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H.,

Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratt, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigó, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.-H., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The Sequence of the Human Genome. *Science*, 291(5507) :1304–1351.

[401] Vingia, S. and Almeida, J. (2003). Alignment-free sequence comparison-a review. *Bioinformatics (Oxford, England)*, 19(4) :513–23.

- [402] Visconti, A., Martin, T. C., and Falchi, M. (2018). YAMP : a containerized workflow enabling reproducibility in metagenomics research. *GigaScience*, 7(7).
- [403] Voelkerding, K. V., Dames, S. A., and Durtschi, J. D. (2009). Next-generation sequencing : from basic research to diagnostics. *Clinical chemistry*, 55(4) :641–58.
- [404] Vogtman, E., Hua, X., Zeller, G., Sunagawa, S., Voigt, A. Y., Hercog, R., Goedert, J. J., Shi, J., Bork, P., and Sinha, R. (2016). Colorectal Cancer and the Human Gut Microbiome : Reproducibility with Whole-Genome Shotgun Sequencing. *PloS one*, 11(5) :e0155362.
- [405] Walsh, R., Thomson, K. L., Ware, J. S., Funke, B. H., Woodley, J., McGuire, K. J., Mazzarotto, F., Blair, E., Seller, A., Taylor, J. C., Minikel, E. V., Exome Aggregation Consortium, MacArthur, D. G., Farrall, M., Cook, S. A., and Watkins, H. (2016). Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genetics in medicine : official journal of the American College of Medical Genetics*.
- [406] Wan, L., Reinert, G., Sun, F., and Waterman, M. S. (2010). Alignment-free sequence comparison (II) : theoretical power of comparison statistics. *Journal of computational biology : a journal of computational molecular cell biology*, 17(11) :1467–1490.
- [407] Watson, J. D. (1990). The human genome project : past, present, and future. *Science (New York, N.Y.)*, 248(4951) :44–9.
- [408] WATSON, J. D. and CRICK, F. H. C. (1953). Molecular Structure of Nucleic Acids : A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356) :737–738.
- [409] Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., Gillespie, J. J., Gough, R., Hix, D., Kenyon, R., MacHi, D., Mao, C., Nordberg, E. K., Olson, R., Overbeek, R., Pusch, G. D., Shukla, M., Schulman, J., Stevens, R. L., Sullivan, D. E., Vonstein, V., Warren, A., Will, R., Wilson, M. J. C., Yoo, H. S., Zhang, C., Zhang, Y., and Sobral, B. W. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Research*, 42(D1) :D581–91.
- [410] Wattam, A. R., Davis, J. J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., Conrad, N., Dietrich, E. M., Disz, T., Gabbard, J. L., Gerdes, S., Henry, C. S., Kenyon, R. W., Machi, D., Mao, C., Nordberg, E. K., Olsen, G. J., Murphy-Olson, D. E., Olson, R., Overbeek, R., Parrello, B., Pusch, G. D., Shukla, M., Vonstein, V., Warren, A., Xia, F., Yoo, H., and Stevens, R. L. (2017). Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Research*, 45(D1) :D535–D542.
- [411] Wayne, L. G., Moore, W. E. C., Stackebrandt, E., Kandler, O., Colwell, R. R., Krichevsky, M. I., Truper, H. G., Murray, R. G. E., Grimont, P. A. D., Brenner, D. J., Starr, M. P., and Moore, L. H. (1987). Report of the Ad Hoc Committee on Reconciliation of

Approaches to Bacterial Systematics. *International Journal of Systematic and Evolutionary Microbiology*, 37(4) :463–464.

- [412] Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., Buck, D., and Au, K. F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, 6 :100.
- [413] Wen, J., Chan, R. H., Yau, S. C., He, R. L., and Yau, S. S. (2014). K-mer natural vector and its application to the phylogenetic analysis of genetic sequences. *Gene*, 546(1) :25–34.
- [414] Wilke, A., Harrison, T., Wilkening, J., Field, D., Glass, E. M., Kyrpides, N., Mavrommatis, K., and Meyer, F. (2012). The M5nr : a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC Bioinformatics*, 13(1) :141.
- [415] Winnenburg, R. (2006). PHI-base : a new database for pathogen host interactions. *Nucleic Acids Research*, 34(90001) :D459–D464.
- [416] Winnenburg, R., Urban, M., Beacham, A., Baldwin, T. K., Holland, S., Lindeberg, M., Hansen, H., Rawlings, C., Hammond-Kosack, K. E., and Kohler, J. (2007). PHI-base update : additions to the pathogen host interaction database. *Nucleic Acids Research*, 36(Database) :D572–D576.
- [417] Winter, S. E., Thiennimitr, P., Winter, M. G., Butler, B. P., Huseby, D. L., Crawford, R. W., Russell, J. M., Bevins, C. L., Adams, L. G., Tsolis, R. M., Roth, J. R., and Bäumlner, A. J. (2010). Gut inflammation provides a respiratory electron acceptor for Salmonella. *Nature*, 467(7314) :426–429.
- [418] Wolever, T. M., Brighenti, F., Royall, D., Jenkins, A. L., and Jenkins, D. J. (1989). Effect of rectal infusion of short chain fatty acids in human subjects. *The American journal of gastroenterology*, 84(9) :1027–33.
- [419] Wong, J. M. W., de Souza, R., Kendall, C. W. C., Emam, A., and Jenkins, D. J. A. (2006). Colonic health : fermentation and short chain fatty acids. *Journal of clinical gastroenterology*, 40(3) :235–43.
- [420] Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, 20(1) :257.
- [421] Wood, D. E. and Salzberg, S. L. (2014). Kraken : ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3) :R46.
- [422] World Health Organization (WHO) (2020). Obesity and overweight (<https://www.who.int/en/news-room/fact-sheets/detail/obesity-and-overweight>).

- [423] Xiao, M., Li, J., Hong, S., Yang, Y., Li, J., Wang, J., Yang, J., Ding, W., and Zhang, L. (2018). K-mer Counting : memory-efficient strategy, parallel computing and field of application for Bioinformatics. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2561–2567. IEEE.
- [424] Xin, H., Lee, D., Hormozdiari, F., Yedkar, S., Mutlu, O., and Alkan, C. (2013). Accelerating read mapping with FastHASH. *BMC genomics*, 14 Suppl 1 :S13.
- [425] Xiong, J., Déraspe, M., Iqbal, N., Krajden, S., Chapman, W., Dewar, K., and Roy, P. H. (2017). Complete Genome of a Panresistant *Pseudomonas aeruginosa* Strain, Isolated from a Patient with Respiratory Failure in a Canadian Community Hospital. *Genome Announcements*, 5(22).
- [426] Xiong, J., Déraspe, M., Iqbal, N., Krajden, S., Chapman, W., Dewar, K., and Roy, P. H. (2017). Complete Genome of a Panresistant *Pseudomonas aeruginosa* Strain, Isolated from a Patient with Respiratory Failure in a Canadian Community Hospital. *Genome announcements*, 5(22).
- [427] Xiong, J., Déraspe, M., Iqbal, N., Ma, J., Jamieson, F. B., Wasserscheid, J., Dewar, K., Hawkey, P. M., and Roy, P. H. (2016). Genome and plasmid analysis of blaIMP-4-carrying *Citrobacter freundii* B38. *Antimicrobial Agents and Chemotherapy*, 60(11) :6719–6725.
- [428] Xu, M., Tao, J., Yang, Y., Tan, S., Liu, H., Jiang, J., Zheng, F., and Wu, B. (2020). Ferroptosis involves in intestinal epithelial cell death in ulcerative colitis. *Cell Death & Disease*, 11(2) :86.
- [429] Yang, J., Chen, L., Sun, L., Yu, J., and Jin, Q. (2008). VFDB 2008 release : an enhanced web-based resource for comparative pathogenomics. *Nucleic acids research*, 36(Database issue) :D539–42.
- [430] Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular biology and evolution*, 10(6) :1396–401.
- [431] Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution*, 11(9) :367–372.
- [432] Ye, S. H., Siddle, K. J., Park, D. J., and Sabeti, P. C. (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell*, 178(4) :779–794.
- [433] Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., and Glöckner, F. O. (2014). The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Research*, 42(D1) :D643–D648.

- [434] Yutin, N. and Galperin, M. Y. (2013). A genomic update on clostridial phylogeny : Gram-negative spore formers and other misplaced clostridia. *Environmental microbiology*, 15(10) :2631–41.
- [435] Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F. M., and Larsen, M. V. (2012). Identification of acquired antimicrobial resistance genes. *The Journal of antimicrobial chemotherapy*, 67(11) :2640–4.
- [436] Zeller, G., Tap, J., Voigt, A. Y., Sunagawa, S., Kultima, J. R., Costea, P. I., Amiot, A., Böhm, J., Brunetti, F., Habermann, N., Hercog, R., Koch, M., Luciani, A., Mende, D. R., Schneider, M. A., Schrotz-King, P., Tournigand, C., Tran Van Nhieu, J., Yamada, T., Zimmermann, J., Benes, V., Kloor, M., Ulrich, C. M., von Knebel Doeberitz, M., Sobhani, I., and Bork, P. (2014). Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular systems biology*, 10 :766.
- [437] Zerbino, D. R. (2010). Using the Velvet de novo Assembler for Short-Read Sequencing Technologies. *Current Protocols in Bioinformatics*, 31(1).
- [438] Zhang, Z., Lin, H., and Ma, B. (2010). ZOOM Lite : next-generation sequencing data mapping and visualization software. *Nucleic acids research*, 38(Web Server issue) :W743–8.
- [439] Zhao, G., Usui, M. L., Lippman, S. I., James, G. A., Stewart, P. S., Fleckman, P., and Olerud, J. E. (2013). Biofilms and Inflammation in Chronic Wounds. *Advances in Wound Care*, 2(7) :389–399.
- [440] Zhou, X. and Li, Y. (2015). Techniques for Oral Microbiology. *Atlas of Oral Microbiology*, pages 15–40.
- [441] Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J., and Wishart, D. S. (2011). PHAST : A Fast Phage Search Tool. *Nucleic Acids Research*, 39(suppl) :W347–W352.
- [442] Zielezinski, A., Vinga, S., Almeida, J., and Karlowski, W. M. (2017). Alignment-free sequence comparison : Benefits, applications, and tools. *Genome Biology*, 18(1) :1–17.
- [443] Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, 29(21) :2669–2677.